基于谱回归和核空间最近邻的基因表达数据分类

于 攀,叶俊勇
(重庆大学光电技术及系统教育部重点实验室,重庆 400030)

摘 要: 肿瘤基因表达数据是典型的高维小样本数据,直接对其进行识别存在维数灾难,需要对数据进行维数约简.提出了一种基于谱回归分析和核空间最近邻分类器的基因表达数据分类方法,采用谱回归分析得到可有效提取低维鉴别特征的投影矩阵,然后通过投影矩阵对基因表达数据进行维数约简,得到的低维数据用核空间最近邻分类器进行识别.通过在 Prostate_Tumor,4_Tumors 两种肿瘤数据集上的实验,证明了该方法的有效性;同时证明了核空间最近邻具有比最近邻更好的分类效果.

关键词: 基因表达数据分类; 核空间最近邻; 谱回归分析; 维数约简
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2011) 08-1955-06

Spectral Regression and Kernel Space K-Nearest Neighbor for Classification of Gene Expression Data

YU Pan, YE Jun-yong
(*Key Laboratory of Optoelectronic Technology and Systems of the Ministry of Education, Chongqing University, Chongqing* 400030, *China*)

Abstract: Cancer gene expression data is a typical data with high dimension and small sample, identifying it directly will encounter the curse of dimensionality, so needs dimensions reduction. This paper proposes a kind of classification approach based on Spectral Regression (SR) analysis and Kernel space K-Nearest Neighbor(KKNN) classifier for gene expression data, it gets the projection matrix through Spectral Regression Analysis witch can extract effectively discriminative characteristics of low dimensions, and reduces the dimensionality of gene expression data by projection matrix, then identifies the low-dimensional data reduced with the Kernel Space K-Nearest Neighbor Classifier. As the experiments operated on the cancer datasets Prostate_Tumor and 4_Tumors demonstrate the effectiveness of the proposed algorithm; simultaneously, compared with the K-Nearest Neighbor(KNN) classification approach, The Kernel space K-Nearest Neighbor has a better classification result.

Key words: gene expression data classification; kernel space k-nearest neighbor; spectral repression analysis; dimensions reduction

1 引言

基因表达数据,或叫微阵列数据,对理解生命起着非常重要的作用[1].从基因数据中挖掘有用的信息还有助于基因的功能研究、基因之间的调控机制及医药研究,并广泛应用于癌症的研究,基于基因表达水平的癌细胞识别为癌症的早期诊断提供了强有力的决策支持,并成为模式识别的热门领域.基因表达数据在数据挖掘分析中,其中最为关键的一步就是进行分类.然而,基因表达数据往往是极少样本上的成千上万个基因的表达.对于基因表达数据的样本分析而言,直接进行样本分类研究时,往往存在维数灾难问题,故有必要对数据进行维数约简[2];同时维数约简也有助于数据的 2 维或 3 维可视化.另外,对于基因表达数据而言,不论是样本分类还是基因分类,直接在基因表达空间进行分类,得到的结果往往不是很理想.究其原因是样本或基因在基因表达空间并不能很好按类别聚类,故为了有效地进行分类识别,需要进行特征提取,即降维.

应用比较广泛的降维算法有主成分分析(Principle Components Analysis, PCA)[3~5],线性判别分析(Linear Discriminant Analysis, LDA)[5~7],保局投影 (Locality Preserving Projection, LPP),近邻保持嵌入(Neighborhood Preserving Embedding, NPE)等.PCA 是在全局最小重构误差的情况下把高维数据投影到低维子空间,而数据点的协

收稿日期:2010-11-25;修回日期:2011-03-02
基金项目:中央高校基本科研业务费(No.10120018)

方差矩阵最大的几个特征值所对应的特征量成为子空间[8],然而 PCA 并没有考虑样本的类别信息,不一定是能很好地区分不同类的样本[9]. LDA 是通过最小化类内散度矩阵和类间散度矩阵的比值来寻找最有效的判别方向[10]. 保局投影方法(LPP)[11]、近邻保持嵌入(NPE)[12]是两种较新的线性投影方法. LPP 通过使用样本最近邻图模型保留样本的近邻结构,在实现降维的同时能很好地保留样本的近邻结构,但是 LPP 并没有考虑样本的类别信息. NPE 同样保留样本局部特性,但是没有考虑到样本的类别信息. 由于它们都需要对稠密矩阵进行特征分解,占用的大量的计算时间和存储空间[13],实验表明这几种降维算法在基因表达数据的识别中误判率相对较高[5,14]. 当样本的特征数量大于样本个数时,LPP 和 NPE 为了获得稳定的最优解,还要进行预处理,如:PCA、SVD,因此实际应用到癌细胞基因表达数据的识别受到很大的限制. 为了克服以上降维算法的缺点,Cai 等人提出了谱回归(Spectral Regression, SR)算法[13],SR 是通过最小化目标函数得到嵌入函数,它保持了数据的整体结构,将流形上的近邻点映射到低维空间时仍为近邻点,远离点映射后仍然为远离点[15]. 相对于前面的几种降维算法,SR 具有更优的计算复杂度和更低的空间复杂度.

本文提出的谱回归和核空间最近邻分类方法是先用 SR 对基因表达数据进行维数约简,然后再用核空间最近邻(Kernel Space K-Nearest Neighbor, KKNN)对降维之后的数据进行分类. 通过谱回归对基因表达数据进行维数约简之后,特征维数降为 c-1[14](c 为肿瘤数据的类别数量),然后再用 KKNN 和 KNN 进行分类. 最近邻(KNN)是一种经典的分类算法,它对低维数据有较好的分类效果,但不是最优的. 本文在 KNN 的基础上提出了一种新的分类算法——核空间最近邻(KKNN),它是通过内积核函数将数据投影到高维空间,然后再利用最近邻算法对高维数据进行分类. 通过用 KKNN 和 KNN 算法分别对 Prostate_Tumor 和 4_Tumors 两个肿瘤基因数据集进行分类实验,表明核空间最近邻的分类效果明显优于最近邻算法.

2 图嵌入的子空间学习算法

本章将从图嵌入的观点出发,为 LDA、LPP、NPE 等子空间学习算法提供一种通用的框架.

对于给定的数据集$\{\boldsymbol{x}_i\}_{i=1}^m$,其中 $\boldsymbol{x}_i \in R^n$,维数约简的目的是在低维空间找到$\{\boldsymbol{z}_i\}_{i=1}^m \subset R^d$, $d << n$. 在图嵌入方法中,给定含有 m 个顶点的图 $\boldsymbol{G}$,每个顶点用一个数据点表示,$\boldsymbol{W}$ 是 $m \times m$ 的对称权值矩阵,$\boldsymbol{W}_{i,j}$为 i, j 的连接权值. 图嵌入的目的是把图中点用低维向量表示,通过连线边权重描述点对间的相似度. 在有监督学习的情况下,假定有 c 类样本,第 t 类有 m_t 个样本,其中 $m_1 + m_2 + \cdots + m_c = m$.

定义 $\boldsymbol{y} = [y_1, y_2, \cdots, y_m]^{\mathrm{T}}$ 是矩阵 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m]$的一维投影,最佳的 $\boldsymbol{y}$ 是在一定约束条件下使下式最小

$$\sum_{i,j} (y_i - y_j)^2 \boldsymbol{W}_{i,j} \tag{1}$$

式(1)可以通过矩阵形式表述

$$\sum_{i,j} (y_i - y_j)^2 \boldsymbol{W}_{i,j} = 2\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{y} = 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{y} \tag{2}$$

式(2)中 $\boldsymbol{D}$ 是对角矩阵,对角线上的元素为 $\boldsymbol{W}$ 的每一行的和(或者每一列的和,因为 $\boldsymbol{W}$ 是对称矩阵),即 $\boldsymbol{D}_{i,i} = \sum_j \boldsymbol{W}_{i,j}$. 加入约束条件 $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y} = 1$,最小问题将转化为求解最优 $\boldsymbol{y}^*$.

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}=1} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{y} = \arg\min_{\boldsymbol{y}} \frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}} \tag{3}$$

将 $\boldsymbol{L} = \boldsymbol{D}$-$\boldsymbol{W}$ 代入式(3)中

$$\begin{aligned}\boldsymbol{y}^* &= \arg\min_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}=1} \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{y} \\ &= \arg\min_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}=1} \frac{\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}} = \arg\min_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}=1} -\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}}\end{aligned} \tag{4}$$

式(4)等效于下式

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}=1} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y} = \arg\max_{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}=1} \frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}} \tag{5}$$

其中最优的 $\boldsymbol{y}$ 是下列特征问题的最大特征向量

$$\boldsymbol{W}\boldsymbol{y} = \lambda\boldsymbol{D}\boldsymbol{y} \tag{6}$$

为了获得所有训练样本和测试样本的映射,选择线性函数 $\boldsymbol{y}_i = f(\boldsymbol{x}_i) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i$,即 $\boldsymbol{y} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$,将其代入等式(5)中

$$\boldsymbol{a}^* = \arg\max_{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}=1} \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}} \tag{7}$$

$\boldsymbol{a}$ 的最优解为下面等式的最大特征向量

$$\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} = \lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} \tag{8}$$

以上是图嵌入的线性扩展(Linear Extension of Graph Embedding, LGE),选择不同的权矩阵 $\boldsymbol{W}$, LGE 将导出很多不同的线性降维算法,如:LDA、LPP、NPE 等,下面将列出与这些算法相对应的的权矩阵 $\boldsymbol{W}$.

LDA:

对于 c 类样本,m_t 表示第 t 类样本的数量,其中 $m_1 + m_2 + \cdots + m_c = m$

$$\boldsymbol{W} = \begin{cases} 1/m_t, & \text{如果 } \boldsymbol{x}_i \text{ 和 } \boldsymbol{x}_j \text{ 同属于第 } t \text{ 类} \\ 0, & \text{其他} \end{cases} \tag{9}$$

很容易得到 $\boldsymbol{D} = \boldsymbol{I}$,是一个单位矩阵.

LPP:

$N_k(\boldsymbol{x}_i)$表示 $\boldsymbol{x}_i$ 的 k 个最近邻点

$$\boldsymbol{W} = \begin{cases} e^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}}, & \text{如果 } \boldsymbol{x}_i \in N_k(\boldsymbol{x}_j) \text{或 } \boldsymbol{x}_j \in N_k(\boldsymbol{x}_i) \\ 0, & \text{其他} \end{cases} \tag{10}$$

其中 $\boldsymbol{D}_{i,i} = \sum_j \boldsymbol{W}_{i,j}$.

NPE:

$N_k(\boldsymbol{x}_i)$表示 $\boldsymbol{x}_i$ 的 k 个最近邻点,$\boldsymbol{M}$ 是一个 $m \times m$ 的重构系数矩阵,$\boldsymbol{M}$ 的定义如下:

对于 $\boldsymbol{M}$ 的第 i 行,如果 $\boldsymbol{x}_j \notin N_k(\boldsymbol{x}_i)$,则 $\boldsymbol{M}_{ij} = 0$;其它情况下,$\boldsymbol{M}_{ij}$由下面的目标函数确定

$$\min \left\| \boldsymbol{x}_i - \sum_{j \in N_k(\boldsymbol{x}_i)} \boldsymbol{M}_{ij} \boldsymbol{x}_j \right\|^2, \sum_{j \in N_k(\boldsymbol{x}_i)} \boldsymbol{M}_{ij} = 1 \tag{11}$$

$\boldsymbol{W}$ 的定义如下

$$\boldsymbol{W} = \boldsymbol{M} + \boldsymbol{M}^{\mathrm{T}} - \boldsymbol{M}^{\mathrm{T}}\boldsymbol{M} \tag{12}$$

很容易证明对角矩阵 $\boldsymbol{D} = \boldsymbol{I}$,即 $\boldsymbol{D}$ 为一个单位矩阵.

3　谱回归降维算法

在实际中,满足 $\boldsymbol{y} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$ 的 $\boldsymbol{a}$ 有可能不存在,最可能的解决办法是用最小二乘算法来最大限度的逼近

$$\begin{aligned} \boldsymbol{a}^* &= \arg\min_{\boldsymbol{a}} \sum_{i=1}^{m} (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{y}_i)^2 \\ &= \arg\min_{\boldsymbol{a}} (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{y})^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{y}) \end{aligned} \tag{13}$$

通过对式(13)的左边求偏导,$\partial(\arg\min_{\boldsymbol{a}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{y})^{\mathrm{T}} \cdot (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{y}))/\partial(\boldsymbol{a}) = 0$,可以得到以下等式

$$\boldsymbol{a} = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{y} \tag{14}$$

当样本的数量小于特征维数时,$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 会是奇异矩阵,等式(14)将呈现病态.一种可能的解决办法是强制对 $\boldsymbol{a}$ 施加一个惩罚因子 α

$$\begin{aligned} \boldsymbol{a}^* &= \arg\min_{\boldsymbol{a}} \left(\sum_{i=1}^{m} (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{y}_i)^2 - \alpha \| \boldsymbol{a} \|^2 \right) \\ &= \arg\min_{\boldsymbol{a}} ((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{y})^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{y}) - \alpha \| \boldsymbol{a} \|^2) \end{aligned} \tag{15}$$

同样将式(15)的右边对 $\boldsymbol{a}$ 求偏导

$$\boldsymbol{a} = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \alpha\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{y} \tag{16}$$

式(16)中,$\boldsymbol{I}$ 是一个 $n \times n$ 的单位阵.很明显,$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \alpha\boldsymbol{I}$ 不再是一个奇异矩阵,$\| \boldsymbol{a} \|^2$ 被称为 Tikhonov 正则化.在统计学上,这种回归算法叫做岭回归.其中当 $\alpha > 0$(本文中 α 取值为 0.01),正则解将不能满足线性方程 $\boldsymbol{y} = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$,且不是特征问题(8)的特征向量;当 α 无限趋近于零时,正则解就是特征问题(8)的特征向量.

对于 c 类的样本,只要具有 c-1 维的特征就能够对其进行分类.所以对于特征问题(16),如果选取最大的 c-1 个特征向量,则最后将得到 c-1 个投影向量,c-1 个投影向量组成一个投影矩阵 $\boldsymbol{A}$,则降维后的样本矩阵 $\boldsymbol{Y} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}$.

图 1 是用谱回归对 4_Tumors 基因表达数据进行降维后的三维分布图,因为有 4 类数据,所以在进行维数约简时,只需要取其中最大的 3 个特征(对于 c 类的样本,只要具有 c-1 维的特征就能对它们进行区分).从图 1 中可以看出 4 类肿瘤基因数据经过谱回归降维之后,能够很清楚的分开,没用重叠区域,说明 SR 能够很好的对基因表达数据进行特征提取.

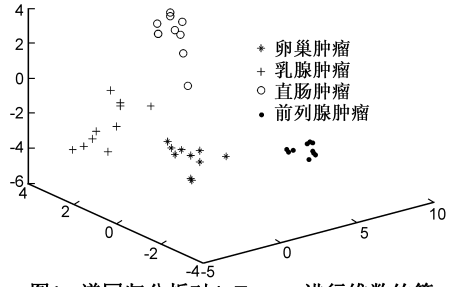

图1　谱回归分析对4_Tumors进行维数约简

4　核空间最近邻分类器

4.1　最近邻分类算法

K 最近邻(K-Nearest Neighbor, KNN)分类算法,是一个理论上比较成熟的方法.该方法的思路是:如果一个样本在特征空间中的 K 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别.该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别.KNN 方法虽然从原理上也依赖于极限定理,但在类别决策时,只与极少量的相邻样本有关.因此,采用这种方法可以较好地避免样本的不平衡问题.另外,由于 KNN 方法主要靠周围有限的邻近的样本,而不是靠判别类域的方法来确定所属类别的,因此对于类域的交叉或重叠较多的待分样本集来说,KNN 方法较其他方法更为适合.

4.2　核空间最近邻分类算法

用谱回归计算得到的投影矩阵 $\boldsymbol{A}$,根据公式 $\boldsymbol{y} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}$计算测试样本 $\boldsymbol{x}$ 投影后的向量 $\boldsymbol{y}$,对降维后的基因数据用最近邻算法进行分类时,发现效果不是很好.因此本文提出了一种基于最近邻的核算法——核空间最近邻算法,它兼有最近邻的稳定性好,抗噪能力强等优点,同时克服了它线性不可分的缺点.

核空间最近邻(KKNN)分类算法的中心思想是,通过一个非线性映射 Φ 将原始特征空间映射到希尔伯特空间 H,在特征空间 H 中再用最近邻分类算法对样本进行分类.

对于降维后的数据集 $\{\boldsymbol{y}_i\}_{i=1}^m$,其中 $\boldsymbol{y}_i \in R^{c-1}$,对应的希尔伯特空间的数据集为 $\{\Phi(\boldsymbol{y}_i)\}_{i=1}^m$,高维空间中点对之间的欧式距离计算公式为

$$\begin{aligned} d(\Phi(\boldsymbol{y}_i), \Phi(\boldsymbol{y}_j)) &= \langle \Phi(\boldsymbol{y}_i) - \Phi(\boldsymbol{y}_j), \Phi(\boldsymbol{y}_i) - \Phi(\boldsymbol{y}_j) \rangle^{1/2} \\ &= (\langle \Phi(\boldsymbol{y}_i), \Phi(\boldsymbol{y}_i) \rangle + \langle \Phi(\boldsymbol{y}_j), \Phi(\boldsymbol{y}_j) \rangle \\ &\quad - 2\langle \Phi(\boldsymbol{y}_i), \Phi(\boldsymbol{y}_j) \rangle)^{1/2} \end{aligned} \tag{17}$$

用核函数 $K(\boldsymbol{x}, \boldsymbol{y})$代替点积 $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle$,式(17)将重写为

$$d(\Phi(\boldsymbol{y}_i),\Phi(\boldsymbol{y}_j))=(K(\boldsymbol{y}_i,\boldsymbol{y}_i)+K(\boldsymbol{y}_j,\boldsymbol{y}_j)-2K(\boldsymbol{y}_i,\boldsymbol{y}_j))^{1/2} \tag{18}$$

由于径向基内积函数是比较常用的核函数,更具有代表性,所以本文中采用径向基内积函数

$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left\{-\frac{|\boldsymbol{x}-\boldsymbol{y}|^2}{256\sigma^2}\right\} \tag{19}$$

σ^2 取值不同会直接影响最后的识别结果,经过多次实验,表 1 所示,当 σ^2 取值为 0.3,识别效果和稳定性最好,因此本文中的 σ^2 都取值为 0.3.

表 1 当 σ^2 取不同值时核空间最近邻对数据集 4_Tumors 的识别率(%)

分类算法	σ^2 的取值					
	$\sigma^2=0.1$	$\sigma^2=0.2$	$\sigma^2=0.3$	$\sigma^2=0.4$	$\sigma^2=0.5$	$\sigma^2=0.6$
KKNN	96.75 (2.25)	96.87 (2.27)	97.12 (2.18)	97.04 (2.95)	96.25 (2.50)	96.12 (2.06)

注:表 1 中括号内的数据为标准差

将待测样本 $\Phi(\boldsymbol{y})$ 与已知类别的训练样本 $\Phi(\boldsymbol{y}_i)$($\Phi(\boldsymbol{y}_i)\in\{\Phi(\boldsymbol{y}_i)\}_{i=1}^m$)计算欧氏距离,并从中选取与 $\Phi(\boldsymbol{y})$距离最小的 k 个点,假设 k_i 是$\Phi(\boldsymbol{y})$的 k 个近邻点中第 i 类的个数($k_1+k_2+\cdots+k_c=k$),然后根据下面的判别式,对 $\Phi(\boldsymbol{y})$进行判别归类

$$k_i=g_i(\Phi(\boldsymbol{y})),\ i=1,2,\cdots,c \tag{20}$$

$g_i(\Phi(\boldsymbol{y}))$为计算 $\Phi(\boldsymbol{y})$的 k 个近邻点中第 i 类的数量的函数,如果

$$k_j=\max\{g_i(\boldsymbol{y})\}_1^c \tag{21}$$

则 $\Phi(\boldsymbol{y})$属于第 j 类,即 $\boldsymbol{y}$ 属于第j 类.

5 实验结果及分析

在这一节中,我们将进行一些实验,以对比本文提出的算法与现存算法之间的优劣.

5.1 实验说明

为了验证本文提出的谱回归和核空间最近邻(SR+KKNN)分类算法的优越性,及在两分类和多分类的情况下的正确率;同时对比谱回归和核空间最近邻分类算法与传统的分类方法的识别效果.本文采用两个具有代表性的肿瘤基因表达数据集 Prostate_Tumor 和 4_Tumors 对其进行测试. Prostate_Tumor 为前列腺肿瘤基因数据,共两类,每类为 50 个样本,共 100 个样本,10509 个特征:一类为肿瘤基因,另一类为正常基因.4_Tumors 共包含 4 类肿瘤基因,每类 20 个样本,共 80 个样本,12533 个特征,4 类肿瘤基因分别为卵巢肿瘤基因、乳腺肿瘤基因、直肠肿瘤基因、前列腺肿瘤基因.

实验步骤如下:

(1)从肿瘤基因数据集中为每类数据随机选取 l (对于 Prostate_Tumors, $l=2,5,10,15,20,25,30$;对于 4_Tumors, $l=4,5,6,7,8,9,10,11,12$)个数据构成训练样本集,其余的基因数据为测试样本集.

(2)分别用 SR、LDA、KDA、PCA、KPCA、LPP、NPE 对训练样本进行特征提取,计算得到一个嵌入函数 $f(\boldsymbol{x})$(即一个投影矩阵 $\boldsymbol{A}$).

(3)然后通过公式 $\boldsymbol{y}=f(\boldsymbol{x})$对测试样本进行降维,对降维之后的测试样本分别用 KKNN 和 KNN 进行分类.对于每类训练样本的 l,为了消除单次选择样本的随机性,独立重复实验 20 次,最后取平均识别率作为最终的识别率,同时得到识别率的方差.

(4)对比同种分类器所对应不同的降维算法的识别率和不同分类器所对应的同种降维算法的识别率,同时通过方差对比 KKNN 和 KNN 的稳定性.

5.2 Prostate_Tumor 数据

各种分类方法在数据集 Prostate_Tumor 上进行测试,相应的分类效果如表 2 所示;同时验证本文提出的谱回归和核空间最近邻分类方法的优越性.

表 2 中,SR+KKNN 的正确率明显高于其他的分类方法,当训练样本数量为 5 时,正确率已达 80%,当训练样本的数量超过 15 时,正确率为 90%以上.随着训练样本的增加其正确率不断的提高,训练样本的数量为 30 个时,正确率接近 93%.对于同样的降维算法,KKNN 的正确率明显高于 KNN,方差却小于 KNN,说明本文提出的 KKNN 分类算法的正确率和稳定性更优.表 2 中 PCA 和 KPCA 的正确率很低,不适于基因表达数据的分类,LDA、KDA、LPP 和 NPE 的效果相对较好,但通过实验数据的对比,表明 SR 的正确率最高的,更适于基因表达数据的分类.

从图 2 中可以直观的看到各种分类器的分类效果,当训练样本数量为 2 时,谱回归和核空间最近邻分类方法的正确率已接近为 69%,高于其他的分类方法.随着样本数量的不断增加,其正确率也逐步升高,并且一直保持着最高的识别率.

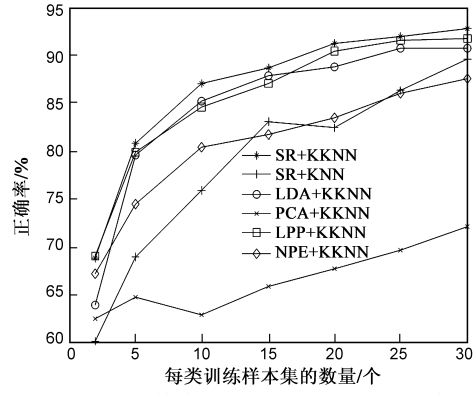

图2 各种分类算法对Prostate_Tumor的识别率曲线

5.3 4_Tumors 数据

各种分类方法在数据集 4_Tumors 上进行测试,相应的分类效果如表 3 所示;同时验证本文提出的谱回归

和核空间最近邻分类方法的优越性.

表 3 中,SR + KKNN 的识别率最高,当训练样本只有 4 个时,其正确率已达到 91%,当训练样本为 11 个时正确率更是接近 98%.而其它的分类方法的正确率相对较低,尤其是用 PCA 和 KPCA 进行降维时,正确率一直低于 80%,根本不适于肿瘤基因表达数据的特征提取.LDA、LPP、NPE 的识别率虽然很高,但是还是略低于 SR + KKNN,且这三种降维算法需要对稠密矩阵进行特征分解,浪费了大量的计算时间和存储空间.对于同一种降维算法,KKNN 的正确率明显高于 KNN,方差却明显小于 KNN,说明相对于 KNN 来说,本文提出的 KKNN 分类算法的正确率和稳定性更优.

表 2　各种分类算法对 Prostate_Tumor 的识别率(%)

分类算法	每类训练样本数量						
	$l=2$	$l=5$	$l=10$	$l=15$	$l=20$	$l=25$	$l=30$
SR + KKNN	**68.85**(9.45)	**80.82**(4.39)	**87.06**(2.98)	**88.71**(4.01)	**91.25**(2.80)	**92.00**(3.00)	**92.75**(2.55)
SR + KNN	60.14(11.49)	68.96(9.22)	75.94(6.88)	83.07(8.37)	82.50(10.67)	86.40(6.24)	89.62(5.92)
LDA + KKNN	63.91(9.87)	79.61(7.06)	85.25(4.47)	87.85(1.93)	88.83(5.09)	90.8(2.86)	90.75(4.26)
LDA + KNN	61.67(7.70)	73.28(11.62)	80.03(7.27)	84.14(7.75)	83.83(8.89)	88.80(2.15)	86.25(6.99)
KDA + KNN	59.47(17.32)	73.77(12.17)	69.55(9.40)	78.5(13.00)	81.15(11.2)	88.40(3.50)	88.75(9.46)
PCA + KKNN	62.50(10.02)	64.78(9.67)	62.94(7.35)	65.93(7.17)	67.75(7.17)	69.70(5.55)	72.13(5.40)
PCA + KNN	60.99(11.11)	64.72(9.48)	62.43(7.11)	64.29(7.22)	65.17(6.52)	68.50(6.38)	72.00(6.27)
KPCA + KNN	58.33(9.75)	60.44(7.32)	64.62(9.44)	64.00(6.32)	60.32(9.22)	63.50(7.33)	64.5(8.88)
LPP + KKNN	68.07(7.46)	79.94(6.42)	84.62(5.18)	87.07(3.08)	90.42(2.08)	91.60(3.53)	91.75(2.74)
LPP + KNN	58.85(8.70)	68.94(10.14)	77.00(9.32)	82.07(10.04)	84.17(7.86)	87.10(6.57)	89.12(4.16)
NPE + KKNN	67.19(7.41)	74.44(7.59)	80.43(3.52)	81.71(5.79)	83.50(4.04)	86.00(4.48)	87.62(4.69)
NPE + KNN	59.79(9.37)	74.39(7.76)	78.69(4.04)	81.36(6.95)	83.25(4.14)	86.00(5.31)	87.12(5.27)

注:表 2 中加粗的数据表示最优的分类方法所对应的识别率,括号内的数据为标准差

表 3　各种分类算法对 4_Tumors 的识别率(%)

分类算法	每类训练样本数量								
	$l=4$	$l=5$	$l=6$	$l=7$	$l=8$	$l=9$	$l=10$	$l=11$	$l=12$
SR + KKNN	**91.01**(4.47)	**92.92**(2.85)	**93.21**(2.75)	**95.58**(3.13)	**95.83**(2.61)	**96.02**(2.54)	**97.13**(1.86)	**97.78**(1.71)	**97.79**(1.68)
SR + KNN	88.98(6.48)	91.83(3.74)	91.87(3.95)	93.94(3.31)	95.31(2.69)	95.23(2.07)	96.75(1.53)	97.36(2.10)	97.47(2.00)
LDA + KKNN	89.84(3.05)	91.17(3.04)	92.50(4.51)	92.89(3.90)	94.58(3.28)	95.23(2.34)	96.00(2.40)	96.94(2.05)	96.98(2.04)
LDA + KNN	87.34(5.58)	88.67(4.43)	91.18(4.05)	90.96(3.63)	94.37(4.51)	95.00(2.26)	95.00(2.26)	96.50(5.31)	96.50(5.31)
KDA + KNN	83.12(3.74)	86.67(7.07)	93.57(4.85)	93.65(3.14)	94.79(1.77)	95.45(2.64)	96.11(2.98)	96.75(1.68)	96.75(1.68)
PCA + KKNN	74.84(5.82)	75.67(6.38)	77.91(5.35)	78.27(5.47)	78.44(5.71)	79.09(5.29)	79.38(7.81)	79.58(4.06)	79.58(4.06)
PCA + KNN	72.93(6.97)	73.58(7.22)	75.37(5.93)	77.40(6.02)	77.92(6.17)	78.52(5.08)	78.75(8.06)	78.06(5.83)	78.06(5.83)
KPCA + KNN	73.12(5.97)	72.33(4.79)	73.21(5.05)	74.92(7.96)	76.13(2.98)	76.32(4.34)	76.50(5.78)	77.22(5.03)	77.22(5.03)
LPP + KKNN	90.08(3.70)	91.08(3.76)	92.95(4.43)	94.13(3.55)	95.72(2.66)	95.90(2.40)	96.50(2.74)	97.22(1.99)	97.33(1.94)
LPP + KNN	86.09(5.19)	89.25(4.54)	91.43(5.45)	92.79(4.08)	94.89(3.20)	95.68(2.85)	95.75(2.94)	96.53(2.01)	96.53(2.01)
NPE + KKNN	88.12(5.17)	91.08(2.97)	91.16(2.68)	91.92(4.26)	92.81(4.13)	92.84(2.88)	93.37(2.95)	94.16(2.97)	94.31(2.91)
NPE + KNN	87.18(5.47)	89.41(4.43)	90.35(4.58)	91.82(5.10)	92.08(4.41)	91.82(3.49)	92.50(3.89)	93.61(3.13)	93.61(3.13)

注:表 3 中加粗的数据表示最优的分类方法所对应的识别率,括号内的数据为标准差

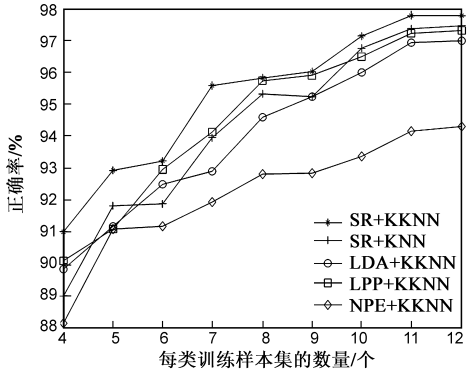

图3　各种分类算法对4_Tumors的识别率曲线

从图 3 可以看出随着训练样本数量的增加,各种分类器的正确率也随之升高,且 SR + KKNN 的识别率一直都高于其它分类算法,当训练样本数量为 4 个时,正确率达到 91%以上.从图 3 中还直观的可以看出 SR + KKNN 的正确率显著高于 SR + KNN,进一步说明 KKNN 的分类效果好于 KNN.

6　结论

大量的实验结果表明,无论是对肿瘤基因和正常基因进行识别的两分类,还是对多种不同肿瘤基因的多分类,SR + KKNN 都表现出比其他分类方法更为优越

的分类效果.对于同一种降维算法,无论是正确率还是稳定性,本文提出的 KKNN 分类算法都明显好于 KNN.通过本文的实验可以看出 PCA 不适于对肿瘤基因数据进行特征提取,LDA、LPP、NPE 虽然效果相对较好,但是稍逊于 SR,并且这三种降维算法由于要对稠密矩阵进行特征分解,占用了大量的计算时间和存储空间,在实际的运用中也收到一定的限制.

参考文献

[1] Baldi P, Hatfield G. DNA Microarrays and Gene Gxpression: from Experiments to Data Analysis and Modeling[M]. Cambridge: Cambridge University Press, 2002. 51 – 84.

[2] L Conde, A Mateo s, 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]. Bourlard H. Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing[C]. New York, USA: IEEE, 2004. 77 – 86.

[3] 赵丽红,孙宇舸,蔡玉. 基于核主成分分析的人脸识别[J]. 东北大学学报(自然科学版),2006,27(8):848 – 850.
Zhao Lihong, Sun Yuge, Cai Yu. Face recognition based on kernel PCA[J]. Journal of Northeastern University(Natural Science), 2006, 27(8): 848 – 850. (in Chinese)

[4] Zhou Jin, Pan Yuqi, Chen Yuehui, Liu Yang. Ensemble classifiers based on kernel PCA for cancer data classification[A]. Huang Deshuang. Emerging Intelligent Computing Technology and Applications: With Aspects of Artificial Intelligence[C]. Berlin, Germany: Springer-Verlag, 2009. 955 – 964.

[5] Tu Chunping, Gan Lan, Yu Zhongping. Based on an improved pre-PCA + LDA classifier design in tumor cells[A]. CCTAE 2010-2010 International Conference on Computer and Communication Technologies in Agriculture Engineering[C]. Piscataway, NJ, USA: IEEE, 2010. 95 – 98.

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

[7] Zhu Lei, Han Bin, Li Hua, Xu Shenhua, Mou Han-zhou, Zheng Zhi-guo. Null space LDA based feature extraction of mass spectrometry data for cancer classification[A]. Wang YQ. Proceedings of the 2009 2nd International Conference on Biomedical Engineering and Informatics[C]. New York, USA: IEEE Medicine and Biology Society, 2009. 1486 – 1489.

[8] P N Belhumeur, J P Hespanha, D J Kriegman. Eigenfaces vs fisherfaces: Recognition using class specific linear projection [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711 – 720.

[9] 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 Yingxin, Ruan Xiaogang. Cancer subtype recognition andfeature selection with gene expression profiles[J]. Acta Electronica Sinica, 2005, 33(4): 651 – 655. (in Chinese)

[10] A M Martinez, A C Kak. PCA versus LDA[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(2): 228 – 233.

[11] He Xiaofei, Cai Deng, Yan Shuicheng, Zhang Hongjiang. Neighborhood preserving embedding[A]. IEEE International Conference on Computer Vision (ICCV'05)[C]. Beijing, China: IEEE, 2005. 1208 – 1213.

[12] He Xiaofei, Cai Deng, Yan Shuicheng, Zhang Hongjiang. Face recognition using laplacianfaces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 328 – 340.

[13] Cai Deng, He Xiaofei, Han Jiawei. Spectral regression for efficient regularized subspace learning[A]. 2007 11th IEEE International Conference on Computer Vision[C]. New York, USA: IEEE, 2007. 214 – 221.

[14] Cai Deng, He Xiaofei, Han Jiawei. SRDA: An efficient algorithm for large scale discriminant analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 1 – 12.

[15] Cai Deng, He Xiaofei, Han Jiawei. Spectral regression: A unified approach for sparse subspace learning[A]. Proceedings of the 7th IEEE International Conference on Data Mining (ICDM'07)[C]. Piscataway, NJ, USA: IEEE Computer Society, 2007. 73 – 82.

作者简介

于 攀 男,1987 年生于江西省丰城市.2009 年毕业于江西理工大学测控技术与仪器专业,获学士学位.现为重庆大学硕士研究生,从事数字图像处理和模式识别方面的有关研究.
E-mail: yupanmail@163.com

叶俊勇 男,1973 年生于四川西昌.现为重庆大学光电工程学院副教授、硕士生导师,主要从事模式识别、图像处理、信号处理等方面的研究工作.
E-mail: ygyocr@cqu.edu.cn