高维少样本数据的特征压缩

游文杰[1,2],吉国力[1],袁明顺[2]
YOU Wen-jie[1,2],JI Guo-li[1],YUAN Ming-shun[2]

1.厦门大学 自动化系,福建 厦门 361005
2.福建师范大学 福清分校,福建 福清 350300
1.Department of Automation,Xiamen University,Xiamen,Fujian 361005,China
2.Fuqing Branch,Fujian Normal University,Fuqing,Fujian 350300,China
E-mail:glji@xmu.edu.cn

YOU Wen-jie,JI Guo-li,YUAN Ming-shun.Feature reduction on high-dimensional small-sample data.Computer Engineering and Applications,2009,45(36):165-169.

Abstract: In view of the characteristics of small sample and high dimensional data,Generalized Small Samples (GSS) is defined. It reduces information feature of GSS:feature extraction (dimensionality extraction) and feature selection (dimensionality selection). Firstly,unsupervised feature extraction based on Principal Component Analysis (PCA) and supervised feature extraction based on Partial Least Squares (PLS) are introduced.Secondly,analyzing the structure of first PC,it presents new global PCA-based and PLS-based feature selection approaches,in addition recursive feature elimination on PLS (PLS-RFE) is realized.Finally,the approaches are applied to the classification of MIT AML/ALL,it performs feature extraction on PCA and PLS,and feature selection compared with PLS-RFE.The information compression of GSS is realized.

Key words: generalized small sample;Principal Component Analysis (PCA);Partial Least Squares (PLS);feature extraction;feature selection

摘 要:针对一类高维少样本数据的特点,给出了广义小样本概念,对广义小样本进行信息特征压缩:特征提取(降维)和特征选择(选维)。首先介绍基于主成分分析(PCA)的无监督与基于偏最小二乘(PLS)的有监督的特征提取方法;其次通过分析第一成分结构,提出基于PCA与PLS的新的全局特征选择方法,并进一步提出基于PLS的递归特征排除法(PLS-RFE);最后针对MIT AML/ALL的分类问题,实现基于PCA与PLS的特征选择和特征提取,以及PLS-RFE特征选择与比较,达到广义小样本信息特征压缩的目的。

关键词:广义小样本;主成分分析(PCA);偏最小二乘(PLS);特征提取;特征选择
DOI:10.3778/j.issn.1002-8331.2009.36.049 **文章编号**:1002-8331(2009)36-0165-05 **文献标识码**:A **中图分类号**:TP391

1 前言

在许多复杂问题中,样本量的绝对数并不算小,但其相对于数据的维数或参数个数而言,样本量就相当小。如20世纪90年代DNA微阵列基因芯片,该技术使得研究人员可以同时测定成千上万个基因的表达水平,得到大量微阵列数据,该数据的特点是样本容量较小,而变量数(基因)非常多。再如,互联网的快速发展,网上出现大量文档数据,自动文本分类也成为处理海量数据的不可或缺的关键技术,其中对使用向量空间模型的分类器的最主要困难是高维的特征空间。这种高维小样本数据对随后的统计分析工作带来了前所未有的困难。

面对这种样本容量小而特征变量数非常多的高维数据,如何建立有效数学模型是一件非常困难的挑战。相对特征变量数而言,这种样本容量数显得非常小的数据,将其定义为广义小样本。所谓广义小样本,是指一类样本容量n远小于其变量维数p,表现为高维数据少样本容量情形。广义小样本是一相对概念,其实质是信息冗余与高噪声,其建模方法的有效性体现在小样本数据潜在信息的充分挖掘,在最大化数据有用信息量的情况下去除冗余与噪声。目前,在数据挖掘中还没有某种方法能普遍适用于各种特点的数据,许多挖掘算法在广义小样本时效率下降甚至失效。构造有效的信息特征压缩方法是广义小样本的一个研究方向。针对广义小样本数据,有两种方法进行信息特征压缩:特征抽取(降维)和特征选择(选维)。

针对高维少样本数据的信息特征压缩问题,为加快特征选择过程,常根据单变量检验统计量的值进行排序(Ranking),如t-检验或信噪比及其p值[1-4],这种操作可能存在一种风险:忽略了特征间相关性及其非线性性。更为精确的方法是要考虑特

基金项目:高校博士点专项科研基金(No.20070384003);福建省教育厅科技项目(No.JB08244)。
作者简介:游文杰(1974-),男,讲师,主要研究方向:统计计算;吉国力(1960-),男,教授,博士生导师,主要研究方向:系统工程理论方法与应用、生物信息学等;袁明顺(1979-),男,硕士,主要研究方向:最优化理论与算法设计。
收稿日期:2009-08-24 修回日期:2009-10-09

征间的联合分布,即同时考虑所有的特征,允许检测那些具有较小主效应,但存在有较强交互效应的特征。该文给出了广义小样本概念,介绍了基于主成分分析(PCA)的无监督特征提取与基于偏最小二乘(PLS)的有监督特征提取两方法;通过分析第一成分结构提出基于PCA与PLS的新的全局特征选择法,借鉴递归特征排除(RFE)[5-6]思想,并进一步提出基于PLS的递归特征排除法(PLS-RFE);最后,在数据集上实现基于PCA与PLS的特征抽取和特征选择,实现广义小样本信息特征压缩。

2 原理方法

2.1 主成分分析(PCA)

PCA是一种重要的无监督特征提取方法。它以较少的潜变量(综合变量)去解释原有数据X中大部分变异,将相关性较强的原变量X转化为互相正交的潜变量T,并从中选取较原变量个数少且能解释大量变异信息的几个新变量(降维),即所谓的主成分。其目标是在低维子空间表示高维数据,使得在误差平方和的意义下低维表示能够最好地描述原始数据。主成分分析是构造原随机变量的一系列线性组合,使各线性组合不相关,且最大可能地包含原变量的信息,即方差最大。

设有n个样本,每一样本观测p个指标:$\boldsymbol{X}=[X_1,X_2,\cdots,X_p]$,$\boldsymbol{X}$的线性组合$\boldsymbol{T}=\boldsymbol{XW}$,使

$$\begin{cases}\max\ var(\boldsymbol{Xw})\\ \text{s.t. } \boldsymbol{w}_i'\boldsymbol{w}_i=1\\ \quad \boldsymbol{w}_i'\boldsymbol{\Sigma}_X\boldsymbol{w}_j=0\\ \quad 1\leqslant i<j\leqslant p\end{cases}$$

称线性组合$\boldsymbol{T}=\boldsymbol{XW}$为主成分,其中$\boldsymbol{\Sigma}_X=\boldsymbol{X}'\boldsymbol{X}$。可以证明[7-8]以上优化问题的解$\boldsymbol{w}_i$满足:

$$(\lambda_i\boldsymbol{I}_p-\boldsymbol{\Sigma}_X)\boldsymbol{w}_i=0$$

$$\boldsymbol{\Sigma}_X=\boldsymbol{X}'\boldsymbol{X}\quad \lambda_i=var(\boldsymbol{t}_i)\quad \lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$$

即优化问题的解$\boldsymbol{w}_i$为$\boldsymbol{\Sigma}_X=\boldsymbol{X}'\boldsymbol{X}$的特征值$\lambda_i$所对应的特征向量$\boldsymbol{w}_i$。也即权重向量$\boldsymbol{W}$(*weighing*)可通过计算协方差阵$\boldsymbol{\Sigma}_X$的特征向量得到,$\lambda_i$表示第$i$个主成分的方差,$\boldsymbol{w}_i$表示第$i$个主轴(*weighing*)。主成分分析的目的之一是简化数据结构,在实际应用中一般选取m($m<p$)个主成分。为确定m取值引进贡献率:

定义1(解释贡献率、累计解释贡献率)

称$\lambda_k/\sum_{i=1}^{p}\lambda_i$为主成分$t_k$的解释贡献率;

称$\sum_{k=1}^{m}\lambda_k/\sum_{i=1}^{p}\lambda_i$为主成分$t_1,t_2,\cdots,t_m$($m<p$)的累计解释贡献率。

累计解释贡献率刻画出m个主成分提取$X_1,X_2,\cdots,X_p$的解释信息量。

2.2 偏最小二乘(PLS)

PLS是一种有监督特征提取方法。它通过主成分分析和综合变量的提取,利用对系统中的数据信息进行分解和筛选,提取对预测变量Y解释性最强的综合变量,辨识系统中的信息与噪声,建立适当的模型。基于PLS的信息特征压缩,在对解释变量数据集X进行压缩的同时,顾及了与预测变量Y的相关程度,其压缩结果将更具有实际意义。

PLS在自变量集X中提取第一潜变量t_1,t_1尽可能多提取X的变异信息;同时在Y中提取第一潜变量u_1使t_1与u_1的相关度最大;建立Y与t_1的回归及X与t_1的回归,若回归方程满足精度要求,则算法结束。否则利用X被t_1解释后的残余信息以及Y被t_1解释后的残余信息进行第二次的潜变量t_2的提取。如此反复,直至达到满足精度要求。

设n个样本p维指标$\boldsymbol{X}=[X_1,X_2,\cdots,X_p]$与预测变量$Y$,其优化模型为:

$$\begin{cases}\max\ cov(\boldsymbol{Xw}_i,\boldsymbol{Yc})\\ \text{s.t. } \boldsymbol{w}_i'\boldsymbol{w}_i=1;\boldsymbol{c}_i'\boldsymbol{c}_i=1\\ \quad \boldsymbol{w}_i'\boldsymbol{\Sigma}_X\boldsymbol{w}_j=0\\ \quad \boldsymbol{c}_i'\boldsymbol{\Sigma}_Y\boldsymbol{c}_j=0\end{cases}$$

其中线性组合$\boldsymbol{t}_i=\boldsymbol{Xw}_i$为第$i$潜变量,$\boldsymbol{\Sigma}_X=\boldsymbol{X}'\boldsymbol{X}$,$\boldsymbol{\Sigma}_Y=\boldsymbol{Y}'\boldsymbol{Y}$。

可以证明[9-10]以上优化问题的解$(\boldsymbol{w}_i,\boldsymbol{c}_i)$为:

$$\boldsymbol{w}_i=\begin{cases}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YX}\text{最大特征值对应特征向量},\ i=1\\ (\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{\Sigma}_{XY}(\boldsymbol{I}-\boldsymbol{P}_Y)\boldsymbol{\Sigma}_{YX}\text{最大特征值对应特征向量},\ i>1\end{cases}$$

$$\boldsymbol{c}_i=\begin{cases}\boldsymbol{\Sigma}_{YX}\boldsymbol{w}_i,\ i=1\\ (\boldsymbol{I}-\boldsymbol{P}_Y)\boldsymbol{\Sigma}_{YX}\boldsymbol{w}_i,\ i>1\end{cases}$$

其中,

$$\boldsymbol{P}_X=(\boldsymbol{\Sigma}_X\boldsymbol{W})[(\boldsymbol{\Sigma}_X\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{\Sigma}_X\boldsymbol{W})]^{-1}(\boldsymbol{\Sigma}_X\boldsymbol{W})^{\mathrm{T}}$$

$$\boldsymbol{P}_Y=(\boldsymbol{\Sigma}_Y\boldsymbol{C})[(\boldsymbol{\Sigma}_Y\boldsymbol{C})^{\mathrm{T}}(\boldsymbol{\Sigma}_Y\boldsymbol{C})]^{-1}(\boldsymbol{\Sigma}_Y\boldsymbol{C})^{\mathrm{T}}$$

$$\boldsymbol{W}=(w_{ij}),\ \boldsymbol{C}=(c_{ij})$$

在PLS计算中所提取成分t_h,一面尽可能多地代表X的变异信息;另一面又尽可能与Y相关联,解释Y中的信息。为测量t_h对X和Y的解释能力,定义t_h的各种解释能力如下。其中$r(x_i,x_j)$表示两变量间相关系数。

定义2(变异解释量、累计变异解释量)定义t_h对X的变异解释能力:

称$Rd(x_j;t_h)=r^2(x_j,t_h)$为成分t_h对自变量x_j的变异解释量;

称$Rd(\boldsymbol{X};t_h)=\frac{1}{p}\sum_{j=1}^{p}Rd(x_j;t_h)$为成分$t_h$对$X$的变异解释量;

称$Rd(\boldsymbol{X};t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}Rd(\boldsymbol{X};t_h)$为成分$t_1,t_2,\cdots,t_m$对$X$的累计变异解释量;

称$Rd(x_j;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}Rd(x_j;t_h)$为成分$t_1,t_2,\cdots,t_m$对$x_j$的累计变异解释量。

同理有,定义t_h对Y的变异解释能力:

称$Rd(y_k;t_h)=r^2(y_k,t_h)$为成分t_h对自变量y_k的变异解释量;

称$Rd(\boldsymbol{Y};t_h)=\frac{1}{q}\sum_{k=1}^{q}Rd(y_k;t_h)$为成分$t_h$对$Y$的变异解释量;

称$Rd(\boldsymbol{Y};t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}Rd(\boldsymbol{Y};t_h)$为成分$t_1,t_2,\cdots,t_m$对$Y$的累计变异解释量;

称$Rd(y_k;t_1,t_2,\cdots,t_m)=\sum_{h=1}^{m}Rd(y_k;t_h)$为成分$t_1,t_2,\cdots,t_m$对$y_k$的累计变异解释量。

3 特征压缩

广义小样本数据的降维压缩方法:特征抽取(降维)和特征选择(选维)。特征提取是将原始的特征空间投影到低维特征空间,投影后的潜在特征是原始特征的线性或者非线性组合,也即特征提取是要对原始的坐标系进行旋转,然后再选取若干重要的潜在特征,显然特征提取是一全局降维方法,当数据集是全局相关时效果较好。特征选择是通过一些标准的统计方法选择出对分类贡献最大的若干特征,它保持原数据主要特征基础

上将数据从高维转成低维,即从原始数据表中选择若干与任务有关的显著特征而构成新的低维数据表,其优点是经特征选择后的数据表没有旋转,其结果易于解释。

3.1 特征提取(降维)

常用的特征提取方法有:

(1) 主成分分析(PCA),它是一种重要的无监督统计分析方法。它能将原始数据空间降维,利用少数几个变量族的线性组合来解释高维变量的协方差结构,挑选最佳潜在特征子集,达到简化数据的目的。

(2) 偏最小二乘法(PLS),它是一种有监督的统计分析方法。它通过主成分分析和综合变量的提取,利用对系统中的数据信息进行分解和筛选,提取对预测变量解释性最强的综合变量,辨识系统中的信息与噪声,建立适当的模型。基于 PLS 的信息特征压缩,在对解释变量数据集进行压缩的同时,顾及了与预测变量的相关程度,其压缩结果将更具有实际意义。

3.1.1 PCA 无监督特征提取

PCA 的特征提取步骤:

步骤 1 标准化数据集,以 $n\times p$ ($p>>n$) 矩阵 X 表示;

步骤 2 计算数据阵 X 的前 m 个主轴 w_i ($i=1,2,\cdots,m$),其中 m 的选取满足 $\sum_{k=1}^{m}\lambda_k/\sum_{i=1}^{p}\lambda_i\geqslant 1-\alpha$,通常 α 取值满足 $1-\alpha\geqslant 0.85$;

步骤 3 计算数据阵 X 在前 m 个主轴 w_i ($i=1,2,\cdots,m$) 上的得分 $T=(t_{ij})=<X_i,w_j>,t_{ij}$ 表示 X_i 在第 j 个主轴上的投影;

步骤 4 得分阵 T 代替原始阵 X 进行相应操作(如判别分类等),性能评价。

3.1.2 PLS 有监督特征提取

PLS 的特征提取步骤:

步骤 1 数据阵 X 以 $n\times p$ ($p>>n$) 表示,编码类别阵 Y 为 $n\times k$ (k 类别数)[3];

步骤 2 计算各成分贡献率及使用"舍一交叉"验证方法,计算预测残差平方和均方(PMPRESS)的最小值对应成分数,及 PMPRESS 对应 $Prob>0.1$ 的最小成分数。同时结合所提取成分对各个变量(自变量与因变量)的解释能力以及累积解释能力,以确定成分数 $nfac$;

步骤 3 计算前 $nfac$ 个成分对应的得分矩阵 $T=(t_{ij})=<X_i,w_j>,t_{ij}$ 表示 X_i 在第 j 个主轴上的投影;

步骤 4 得分阵 T 代替原始阵 X 进行相应操作(如判别分类等),性能评价。

3.2 特征选择(选维)

广义小样本问题的一个实际任务是:用最少的特征变量实现最优的目标(如最大识别率)。也即选择数量少而携带信息量大的特征变量,一方面能最大地去除冗余与噪音,另一方面能大量减少实际操作成本。特征选择通常分为两个阶段,首先基于 Filter 方法从成千上万的特征中筛选出一定量的特征,以降低搜索空间,其次基于 Wrapper 方法进一步选出满足条件的显著特征子集。如何从众多特征中寻找一组最有效特征是问题的关键,以下提出基于 PCA 与 PLS 的新的全局特征选择方法,及基于 PLS 的递归特征排除法(PLS-RFE)。

3.2.1 PCA 特征选择

由 2.1 节的分析,可得以下结论:设 t_1 是 X 的第一主成分,则 t_1 与原始数据阵 X 的综合相关度最大,即 $\sum_{j=1}^{p}\rho^2(t_1,X_j)=\lambda_1$ 最大。也即若只选取一个综合变量代替原始变量 X,则 t_1 是 X 的最优选择。第一成分 t_1 对应于数据变异最大的方向,即 t_1 是使数据信息损失最小、精度最高的一维综合变量。所以从 w_1 系数的大小、符号上分析:系数绝对值较大,则表明该主成分主要综合了绝对值大的特征变量,正号表示变量与主成分作用同向,负号表示原变量与主成分作用反向。若只选取第一成分,则从 w_1 系数中选择分量绝对值大的特征变量,实现基于 PCA 的特征选择。

3.2.2 PLS 特征选择

同理,由 2.2 节的分析知,PLS 建模中要求:(1) t_1 和 u_1 各自提取 X 与 Y 中尽可能多的变异信息;(2) t_1 和 u_1 的相关性达到最大。也即若只选取一个潜变量代替原始变量 X,则 t_1 是 X 的最优选择。第一成分 t_1 对应于数据集 X 变异尽可能大的方向,即 t_1 是使数据集 X 信息损失尽可能小、精度尽可能高的一维综合潜变量。所以从 w_1 系数的大小分析:系数绝对值较大,则表明该成分主要综合了绝对值大的特征变量。若只选取第一成分,则从 w_1 系数中选择分量绝对值大的特征变量,实现基于 PLS 的特征选择。

3.2.3 PLS-RFE 特征选择

实际问题中,通常只有少量的特征是真正的与目标信息(如类别)相关,而大部分特征是与目标信息无关的"噪音"。在对目标信息进行分析时,过多的"噪音"特征将干扰有用信息,使计算出来的特征权值失真,影响特征排序的准确性。这里借鉴递归特征排除(RFE)思想,提出基于 PLS 的递归特征排除法 PLS-RFE (Recursive Feature Elimination),其步骤:(1) 对特征集中的所有特征由 3.2.2 节中的 PLS 方法进行特征重要性排序(Feature Ranking),删除排列最后的特征;(2) 余下特征重新由 PLS 方法计算,再删除排列最后的特征,如此反复,直至保留特征集中的 k 个特征,实现基于 PLS-RFE 的特征选择。

4 实验分析

4.1 数据

急性白血病是儿童肿瘤中发病率占第一位的疾病,在临床上,根据白血病细胞的形态及组织化学染色表现,可将此病分为急性淋巴细胞性白血病(Acute Lymphoblastic Leukemia,ALL)以及急性髓细胞性白血病(Acute Myeloid Leukemia,AML)两大类。急性白血病不论何种细胞类型,其主要临床表现大致相似,且白血病的初期症状可能不明显,与一般常见儿童疾病症状类似。所以对急性淋巴细胞性白血病与急性髓细胞性白血病的准确识别,对急性白血病的早期诊断和针对性治疗以及提高生存率和生存质量都有很大的帮助。美国麻省理工学院的 Golub[1]等人使用高密度寡核苷酸阵列检测了 7 129 个基因表达水平,原始训练数据包含 38 个样本(27 个 ALL,11 个 AML);测试数据包含 34 个样本(20 个 ALL,14 个 AML)。Golub 等人筛出 50 个基因,并根据 38 个训练样本构造了一个分类器,应用于 34 个新收集到的测试样本上,结果有 29 个样本被正确识别。

4.2 实验

这里选择支持向量机(SVMs)作为分类器,基于 Matlab 平台的 SVMs 工具箱 OSU_SVM3.00,下载地址:http://www.kernel-methods.net/。选择线性核函数 LinearSVC,相应参数取默认值。首先,对数据集的所有特征分别采用基于PCA/PLS 特征提取

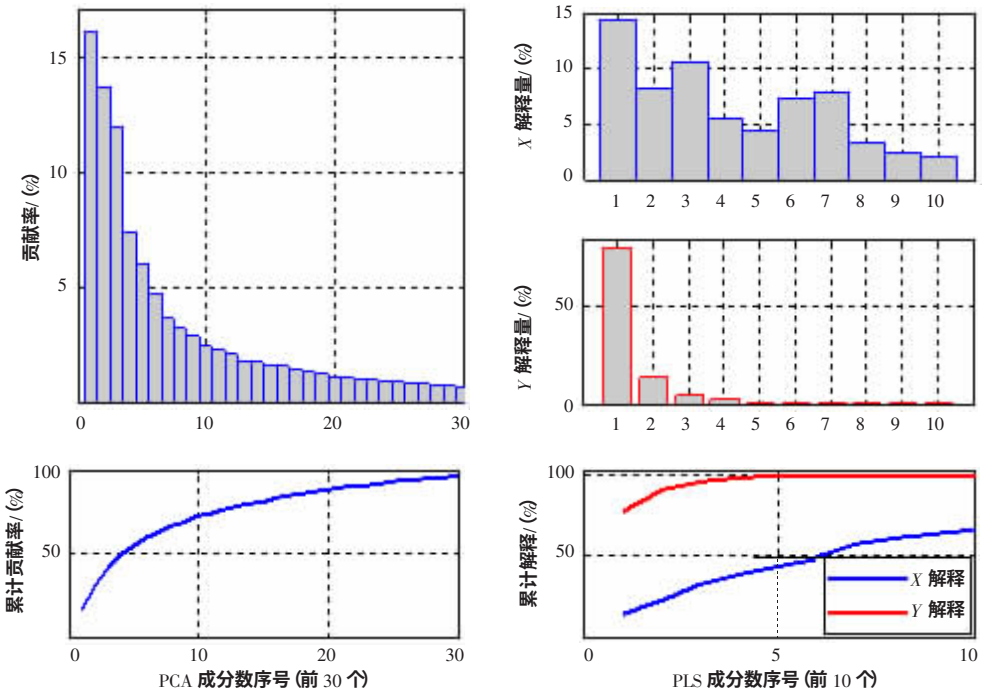


图 1 基于 PCA/PLS 的(累计)贡献/解释与成分数之间的关系(训练集)

(特征选择)选择前 k(k=2,3,…,10)个综合特征(信息特征);其次,将所选取的综合特征(信息特征)进行支持向量机(SVMs)分类训练;最后,分别在训练样本与测试样本上进行测试,计算识别率,并进行校验分析。

4.2.1 降维

分别使用 PCA 与 PLS 方法来进行特征提取,并对所提取"潜变量"进行比较分析。步骤为:

(1)使用 PCA(PLS)对数据集进行降维,以 7 129 个基因表达水平为原始数据空间;

(2)结合各成分贡献率(图 1)及 SVMs 正确识别率,选择恰当的"综合特征"数。

表 1 为全部(7 129 个)特征经特征提取后的前 10 个"潜变量"的 SVMs 识别结果:

表 1 基于 PCA/PLS 的特征提取所选前 10 个"潜变量"的识别率

成分数	基于 PCA 的识别率			基于 PLS 的识别率		
	训练集	测试集	支持向量	训练集	测试集	支持向量
2	**1.000 0**	**0.882 4**	**(1,2)**	**0.868 4**	**0.970 6**	**(12,11)**
3	1.000 0	0.852 9	(2,2)	1.000 0	0.911 8	(12,11)
4	1.000 0	0.852 9	(2,3)	1.000 0	0.882 4	(13,11)
5	1.000 0	0.823 5	(4,2)	1.000 0	0.911 8	(13,11)
6	1.000 0	0.852 9	(2,4)	1.000 0	0.911 8	(13,11)
7	1.000 0	0.852 9	(3,3)	1.000 0	0.911 8	(13,11)
8	1.000 0	0.705 9	(4,3)	1.000 0	0.882 4	(14,11)
9	1.000 0	0.764 7	(3,4)	1.000 0	0.882 4	(16,11)
10	1.000 0	0.764 7	(4,4)	1.000 0	0.882 4	(17,11)

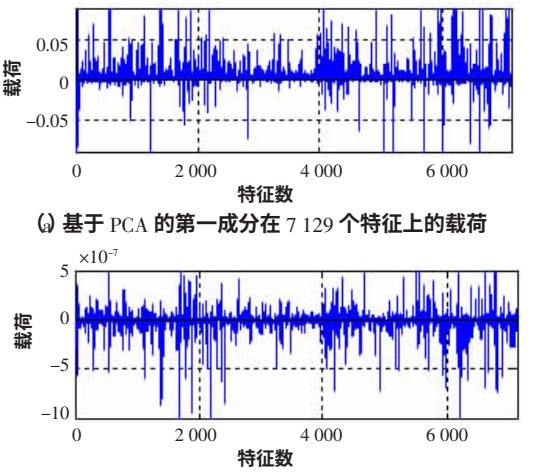
注 数据集:MIT AML/ALL,分类器:SVMs(OSU_SVM3.00),线性核,参数默认。

经 PCA 与 PLS 特征提取后的前 k 个"综合特征"在 SVMs 分类器的识别率如表 1,在成分数为 2 时识别率最高。经 PCA 特征提取后的训练集与测试集识别率分别为 100%与 88.24%,而经 PLS 特征提取后的训练集与测试集识别率分别为 86.84%与 97.06%。这结论符合 Nguyen[2-4]等提出的直接选取前 3 个综合特征的做法。并且当成分数增加时,基于 PLS 的测试样本识别率明显优于 PCA 的识别率。

从图 1 中知,成分数为 2 时所对应 PCA 的变量变异解释原始信息达到 30%,对应于 PLS 对原变量变异的解释:解释自变量变异 23%,同时解释因变量 91%的信息。结合表 1 结论可知,在众多特征(7 129 个)中只有少量的特征是真正的与样本类别相关,而大部分特征是与样本类别无关的"噪音"。在图 1 中明显看出基于 PCA/PLS 第一成分所携带的信息量最大,故可以第一成分所刻画的权值进行特征选择。

4.2.2 选维

由 2.2 节的分析,第一成分携带原数据变异信息最大,所以从第一成分权值(载荷)进行特征选择,系数绝对值较大,则表明该特征在解释第一成分时更重要,也即在解释原数据时贡献大,如图 2。



(a)基于 PCA 的第一成分在 7 129 个特征上的载荷

(b)基于 PLS 的第一成分在 7 129 个特征上的载荷

图 2 第一成分上的载荷与特征变量之间的关系

以下就以基于 PCA/PLS 的第一成分权值进行特征选择。具体步骤为:

(1)特征选择:基于 PCA/PLS/PLS-RFE 的特征选择方法进

表 2 基于 PCA/PLS 的特征选择所选前 k 个特征的识别率

特征数	基于 PCA 的特征选择			基于 PLS 的特征选择			基于 PLS-RFE 的特征选择		
	训练集/(%)	测试集/(%)	支持向量	训练集/(%)	测试集/(%)	支持向量	训练集/(%)	测试集/(%)	支持向量
2	71.1	58.8	(11,10)	84.2	85.3	(4,5)	84.2	85.3	(4,5)
3	86.8	73.5	(5,3)	89.5	76.5	(4,5)	89.5	76.5	(4,5)
4	86.8	88.2	(4,4)	100	94.1	(2,3)	100	94.1	(2,3)
5	86.8	85.3	(4,3)	100	79.4	(3,3)	100	79.4	(3,3)
6	81.6	47.1	(8,2)	100	94.1	(3,3)	100	94.1	(3,3)
7	92.1	61.8	(7,3)	100	91.2	(4,3)	100	91.2	(4,3)
8	92.1	79.4	(7,4)	100	91.2	(5,2)	100	91.2	(5,2)
9	100	85.3	(4,3)	**100**	**100**	**(5,3)**	**100**	**100**	**(5,3)**
10	100	88.2	(5,3)	100	85.3	(7,2)	100	85.3	(7,2)
11	100	85.3	(4,3)	100	85.3	(6,3)	100	85.3	(6,3)
12	100	73.5	(4,4)	100	82.4	(7,2)	100	82.4	(7,2)
13	**100**	**91.2**	(4,5)	100	88.2	(8,2)	100	88.2	(8,2)
14	100	79.4	(5,5)	100	91.2	(7,2)	100	91.2	(7,2)
15	100	79.4	(5,5)	100	88.2	(6,3)	100	88.2	(6,3)
7 129	100	97.1	(15,7)	100	97.1	(15,7)	100	97.1	(15,7)

注 数据集:MIT AML/ALL,分类器:SVMs(OSU_SVM3.00),线性核,参数默认。

行特征筛选。选择前 k (k=2,3,…,15) 个特征。

(2) 分类器:以支持向量机为分类器进行分类,选择线性核函数 LinearSVC,相应参数取默认值。

(3) 计算识别率:分别在训练样本与测试样本上进行测试,计算识别率。结果如表 2。

相比较于表 1,显然在特征选择后 PLS 与 PLS-RFE 的识别率已达到 100%,也即在去除冗余与噪声后,分类器 SVMs 表现更优。同时,从表 2 知基于 PCA 在选择 13 个特征时训练集全部识别,测试集识别达到 91.2%;而基于 PLS 与 PLS-RFE 在选择 9 个与 9 个特征时训练集与测试集均全部正确识别。PLS 与 PLS-RFE 方法的结果明显好于 Golub 等人的结果。

4.3 评价

由 SVMs 基于数据集 MIT AML/ALL 进行特征选择与分类,分别采用留一校验(LOOCV)算法、k-折叉校验(k-fold CV)算法和保留法(holdou)来评价文中的方法。结果如表 3,其中在 k 折叉法(4-fold)进行 PLS-RFE 特征选择,平均选择 6.41 个特征时训练与测试均 100%识别,结果好于 PLS 的结果。在留一校验(LOOCV)法中,不论是 PLS 还是 PLS-RFE 均出现一个错分 #66,这在 Golub[1]等人的工作中同样错分了此样本,甚至有人[3]认为这些样本可能存在错误标记。

表 3 实验评价结果

特征选择算法	校验方法	(平均)选择特征数	误判数	备注
PLS	留一法(72 个样本)	5.01	1	误判样本#66
	k 折叉法(4-fold)	6.95	0	随机 50 次
	保留法(训练 38 个测试 34 个)	9	0	表 2
PLS-RFE	留一法(72 个样本)	5.01	1	误判样本 #66
	k 折叉法(4-fold)	6.41	0	随机 50 次
	保留法(训练 38 个测试 34 个)	9	0	表 2

注 训练与测试集 100%识别时,平均选择最少的特征数。

5 总结

在高维少样本数据的压缩中,PCA 能有效概括原数据的结构特征,其优点是数据压缩充分,生成综合特征数少。但其不足在于所选取主成分与预测变量 Y 无关,只针对解释变量 X 去寻找对其解释重要的成分,与预测变量 Y 相关性大却在解释变量 X 中所占比例小的成分有可能被删除。而 PLS 克服了这些不足,其在对解释变量 X 进行压缩时,顾及与预测变量 Y 的相关程度。

文章对主成分降维和偏最小二乘降维进行讨论,并尝试利用主成分和偏最小二乘进行选维操作,提出基于 PCA 与 PLS 的特征选择及 PLS-RFE 特征选择方法。并针对目前常用的基于单变量检验统计量的特征选择存在的不足,提出基于 PCA 与 PLS 的新的全局特征选择法,并借鉴递归特征排除(RFE)思想,提出基于 PLS 的递归特征排除法(PLS-RFE);最后在数据集上实现基于 PCA 与 PLS 的特征抽取和特征选择,实现广义小样本信息特征压缩。

参考文献:

[1] Golub T R,Slonim D K,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] Nguyen D V,Rocke D M.Tumor classification by partial least squares using microarray gene expression data[J].Bioinformatics,2002,18(1):39-50.

[3] Nguyen D V,Rocke D M.Multi-class cancer classification via partial least squares with gene expression profiles[J].Bioinformatics,2002,18(9):1216-1226.

[4] Nguyen D V,Rocke D M.On partial least squares dimension reduction for microarray-based classification:A simulation study[J].Computational Statistics & Data Analysis,2004,46(9):407-425.

[5] Guyon I,Weston J,Barnhill S,et al.Gene selection for cancer classification using support vector machines[J].Machine Learning,2000,46(13):389-422.

[6] 阮晓钢,李颖新,李建更,等.基于基因表达谱的肿瘤特异基因表达模式研究[J].中国科学:C 辑,2006,36(1):86-96.

[7] 高惠璇.应用多元统计分析[M].北京:北京大学出版社,2005:265-277.

[8] Massey W F.Principal components regression in exploratory statistical research[J].Journal of American Statistical Association,1965,60:234-246.

[9] Wold S,Ruhe A,Wold H,et al.The collinearity problem in linear regression,the partial least squares (PLS) approach to generalized inverses[J].Journal of Statistics Computation,1984,5:735-743.

[10] Lorber A,Wangen L,Kowalski B.A theoretical foundation for the PLS algorithm[J].Journal of Chemometrics,1987,1:19-31.