Statistics and Application 统计学与应用, 2019, 8(5), 835-841
Published Online October 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.85094

Hans汉斯

Application of Partial Least Squares Method in Diagnosis of Esophageal Squamous Cell Carcinoma (ESCC)

Yuguo Li, Xiaotong Li*, Yanwen Yan, Shen Fan

Department of Mathematics, College of Science, China University of Petroleum (Beijing), Beijing
Email: *fly6688@126.com

Received: Oct. 8th, 2019; accepted: Oct. 24th, 2019; published: Oct. 31st, 2019

Abstract

In this paper, partial least squares method and principal component method are used to reduce the dimension of serum microRNAs data associated with Esophageal Squamous Cell Carcinoma (ESCC), and several traditional discriminant methods and 10-fold cross-validation are employed to reduce the dimension of the data, and a higher classification accuracy is obtained. By comparing the accuracy of the two methods, it shows that the partial least squares method performs better than the principal component method in dimension reduction and the reasons are analyzed.

Keywords

Partial Least Squares, Dimension Reduction, Discriminant Analysis, ESCC

偏最小二乘法在食管鳞癌诊断中的应用

李玉国，李晓童*，严彦文，范　申

中国石油大学(北京)理学院数学系，北京
Email: *fly6688@126.com

收稿日期：2019年10月8日；录用日期：2019年10月24日；发布日期：2019年10月31日

摘　要

本文采用偏最小二乘法和主成分法分别对与食管鳞癌相关血清miRNAs数据进行降维处理，对降维数据利用传统的几种判别方法并结合十折交叉验证，得到了较高的分类准确率，对比两种方法的准确率，表明了在降维方面偏最小二乘法比主成法表现更好并分析了原因。

*通讯作者。

文章引用: 李玉国, 李晓童, 严彦文, 范申. 偏最小二乘法在食管鳞癌诊断中的应用[J]. 统计学与应用, 2019, 8(5): 835-841. DOI: 10.12677/sa.2019.85094

关键词

偏最小二乘法，降维，判别分析，食管鳞癌

1. 引言

基于基因表达血清 miRNAs 数据或单核苷酸多态性(SNPs)数据，进行疾病的诊断，是一种被广泛应用的医学诊断方法。在食管鳞癌的诊断方面，传统的诊断方法有内镜检查和 X 线检查等方法。但是，前者具有侵入性，易引起患者的不适，使得大部分人的接受度较差，而且最早阶段性食管鳞癌患者无症状，其病变局限于粘膜或粘膜下层，因此通常会失去被早期诊断的机会。此外，侵袭性和样本误差的可能性限制了内镜活检的有效性[1]。后者需要使用钡餐，然后照射 X 射线，对身体具有潜在的放射性危险。而基于血清 miRNAs 的方法，规避了以上方法的缺点，简单易于操作，对于患者产生的不适性较小，患者的接受程度较高，因而被广泛地应用于医学肿瘤诊断方面，进而成为生物医学统计的一个热点。

但是，在通常情况下这类数据包含的样本的容量远远小于样本的维数，且数据中的变量之间存在严重的多重共线性。这使得利用传统的统计学方法得到的分类效果较差。针对这种情况，学者们提出了各种降维的方法，总的归纳起来可以主要分为两类，一类是带有惩罚项的变量选择，如 R. Tibshirani [2]和 Y. Liu [3]带有 L_1 范数惩罚项的 LASSO 法和带有 L_2 范数惩罚项的“岭回归”法，或者联合多种变量选择的方法共同使用[1]。第二类方法通过原始变量的线性组合来实现降维，如主成分(PCA)降维法和偏最小二乘(PLS)降维法。两种方法各有优势。

Barker M 和 Rayens W [4]从理论上证明了源于计量化学领域回归问题的以解决多重共线性为目的偏最小二乘法与 Fisher 线性判别法之间的关系。本文采取了基于偏最小二乘(PLS)和主成分(PCA)的降维方法，分别对于与食管鳞癌相关的血清 miRNAs 数据降维处理，并进一步采用线性判别分析(LDA)，二次判别分析(QDA)，逻辑斯蒂判别(Log)，支持向量机(SVM)等判别方法，把数据分为训练集和测试集，利用十折交叉验证的方法，在测试集上得到了较高的分类准确率。

2. 数据与方法

2.1. 数据来源

国家生物信息中心(NCBI，编码：GSE112840)。我国是食管鳞癌发病率和死亡率最高的国家，尤其是位于太行山脉南部地区的河南省林州市等地是我国食管鳞癌的高发区。该疾病产生的主要原因是患者长期不良的生活习惯和疾病史。数据中的所有食管鳞癌病例均为首次病理诊断，于 2014 年 10 月至 2015 年 10 月在河南省林州市肿瘤医院内镜中心连续招募[1]。本文中与食管鳞癌相关血清 miRNAs 数据，共包含共 104 个样本(其中 52 个是健康者，52 个为患者)。本文采用该数据的 1918 个特征(从 Blank 到 Negative Control)，作为判断是否患病的判别依据，使用哑变量来表示类别，其中 1 表示健康，0 表示患病。

2.2. 方法

本文采用三阶段法，1) 基因选择。先从原始的 1918 个特征中利用 t-统计量的方法，选取 m 个特征。

2) 降维。在这 m 个特征基础上，利用偏最小二乘法和主成分分析法降维到三维。3) 判别。在三维数据的基础上，利用线性判别法(LDA)，二次判别法(QDA)，逻辑斯蒂判别法(Log)和支持向量机(SVM)等常用的统计方法实现对数据的分类判别。

2.2.1. 基因选择

原始的数据属于高维数据，在众多的特征之中，存在大量的对于分类无用的噪音，如何剔除掉这些无用的噪音进而筛选出对于判别“贡献率”大的特征对于判别的准确率至关重要。因此引入如下定义的简单 t-统计量[5]：

$$t=\frac{\overline{x}_0-\overline{x}_1}{\sqrt{\frac{s_0^2}{N_0}+\frac{s_1^2}{N_1}}}$$

其中 $N_k,s_k^2,\overline{x}_k$ $(k=0,1)$分别表示每一类的数量，方差和均值。计算每一个特征的简单 t-统计量，然后选取简单 t-统计量为正的最大的前 $m/2$ 和 t-统计量为负的最小的后 $m/2$ 作为第一阶段选取的 m 个特征。在本文中 m 取值分别为 20 和 50。

2.2.2. 降维方法

1) 偏最小二乘降维法[4]

偏最小二乘法是源于计量化学领域的一种高效率的多元分析统计分析方法。如今，已经广泛地应用于生物医学，基因分析领域。其定义如下：设自变量 $x_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)'$，相应的自变量矩阵为：$X=\left(x_{ij}\right)_{n\times p}$，$i=1,2,\cdots,n,j=1,2,\cdots,p$。响应变量为：$y_i=\left(y_{i1},y_{i2},\cdots,y_{iq}\right)'$。对应的响应变量矩阵为：$Y=\left(y_{ij}\right)_{n\times q}$，$i=1,2,\cdots,n,j=1,2,\cdots,q$。然后分别寻找自变量和响应变量的线性组合，使得这两个线性组合之间的相关性达到最大。即：

$$\arg\max\left\{\frac{\left[\operatorname{cov}\left(a^{\mathrm{T}}x,b^{\mathrm{T}}y\right)\right]^2}{\left(a^{\mathrm{T}}a\right)\left(b^{\mathrm{T}}b\right)}\right\}=\{a\} \tag{1}$$

a 为矩阵 $\Sigma_{xy}\Sigma_{yx}$ 的最大的特征值所对应的特征向量。其中 Σ_{xy} 为自变量与响应变量的协方差矩阵。把 a 成为响应变量与自变量之间的最大偏最小二乘方向。

引理[6]：设 B 是 P 阶对称矩阵，λ_i 是 B 的第 i 大特征值，l_i 是相应于 λ_i 的 B 的标准化特征向量 $(i=1,2,\cdots,p)$，x 为任意一个非零的 p 维向量，那么有：

$$\lambda_p\le\frac{x'Bx}{x'x}\le\lambda_1$$

上式右边等号当 $x=cl_1$ 时成立，左边等号当 $x=cl_p$ 时成立，这里 c 是常数。

下面证明当 a 为矩阵 $\Sigma_{xy}\Sigma_{yx}$ 的最大的特征值所对应的特征向量时(1)取得最大值。

$$\arg\max\left\{\frac{\left[\operatorname{cov}\left(a^{\mathrm{T}}x,b^{\mathrm{T}}y\right)\right]^2}{\left(a^{\mathrm{T}}a\right)\left(b^{\mathrm{T}}b\right)}\right\}=\arg\max\left\{\frac{\left(a^{\mathrm{T}}\Sigma_{xy}b\right)^2}{\left(a^{\mathrm{T}}a\right)\left(b^{\mathrm{T}}b\right)}\right\}=\arg\max\left\{\frac{1}{\left(a^{\mathrm{T}}a\right)}\frac{\left(b^{\mathrm{T}}\Sigma_{yx}a\right)\left(a^{\mathrm{T}}\Sigma_{xy}b\right)}{\left(b^{\mathrm{T}}b\right)}\right\}$$

$$=\arg\max\left\{\frac{a^{\mathrm{T}}\Sigma_{xy}\Sigma_{yx}a}{\left(a^{\mathrm{T}}a\right)}\right\}$$

a 为矩阵 $\Sigma_{xy}\Sigma_{yx}$ 的最大的特征值所对应的特征向量， $b=\Sigma_{yx}a$ 。

偏最小二乘降维法如下： $X_0=X, Y_0=Y, h=1,2,3$

1) $\arg\max\left\{\dfrac{\left[\operatorname{cov}\left(a_h^{\mathrm{T}}X_{h-1}, b_h^{\mathrm{T}}Y_{h-1}\right)\right]^2}{\left(a_h^{\mathrm{T}}a_h\right)\left(b_h^{\mathrm{T}}b_h\right)}\right\}=a_h$；

2) $\xi_h=X_{h-1}a_h$；

3) $t_h=X_{h-1}^{\mathrm{T}}\xi_h\Big/\left(\xi_h^{\mathrm{T}}\xi_h\right)$；

4) $X_h=X_{h-1}-\xi_h t_h^{\mathrm{T}}$；

5) $X\to\left(\xi_1,\xi_2,\xi_3\right)$。

b) 主成分降维法[6]

主成分降维法是多元统计分析中常用的一种方法。找到自变量的线性组合，使得自变量在该线性组合下的方差达到最大。定义如下：

$$\arg\max_{a'a=1}\left[\operatorname{cov}\left(a^{\mathrm{T}}x\right)\right]=a_k$$

已知 a_k 为矩阵 Σ_{xx} 的从大到小的第 k 个特征值所对应的特征向量。其中 Σ_{xx} 为自变量与响应变量的协方差矩阵。我们采用前三个最大的特征值所对应的特征向量，把原始数据集 X 投影到这三个方向上，将维数降低到三维。

2.3. 判别方法

本文采用的判别方法有逻辑斯蒂判别(Log)，线性判别(LDA)，二次判别(QDA)和支持向量机(SVM)。以上方法的定义均可在参考文献[7]中查到，在此不再赘述。

3. 结果及其分析

我们使用十折交叉验证(10-fold cross-validation)的方法，将数据集分成十份，轮流将其中 9 份作为训练数据集，1 份作为测试数据集，进行试验。以在测试集上的错误率作为判断该方法优劣的一个标准。经过 100 次十折交叉验证后的平均错误率如下表 1 所示：

Table 1. Comparisons of classification error rates of four discrimination methods under two dimension reduction methods. Classification results of serum microRNAs data in training set and test set. Each value in the table is error classification percentage averaged over 100 times of 10-fold cross validation. Perfect error classification is 0.1675

表 1. 在两种降维法下四种判别方法的错判率比较

m	Log		LDA		QDA		SVM	
	PLS	PCA	PLS	PCA	PLS	PCA	PLS	PCA
20	0.1773	0.2797	0.1744	0.2763	0.3546	0.2840	0.1685*	0.2648#
	0.1311	0.2599	0.1290	0.2600	0.1289	0.2274	0.1291*	0.2586#
50	0.1803	0.2588	0.1887	0.2576	0.4091	0.3078	0.1675**	0.2484##
	0.1279	0.2354	0.1452	0.2432	0.1457	0.2563	0.1507**	0.2293##

注：*cost = 10，#cost = 0.01，**cost = 0.005，##cost = 1。

其中第一行与第三行表示在当 $m = 20$ 和 $m = 50$ 时测试集上的平均错误率。第二行与第四行表示当 $m = 20$ 和 $m = 50$ 时在训练集上的平均错误率(见表 1)。显然，测试集上的平均错误率一致地大于训练集上的错误率，这与我们通常的结论是一致的。另外，如表 1 所示，我们也可以明显地看出，无论是在训练集上还是测试集上，利用偏最小二乘降维得到的平均错误率一致地小于由主成分降维得到的平均错误率。大概低于 10%以上。在充分的证明了偏最小二乘法在降维上相对于一般的主成分法的巨大优势。究其原因，在于以下两点：1) 参考文献[4]证明了，在二元判别中，偏最小二乘的第一个降维方向与费舍尔(Fisher)的线性判别方向是一致的。即考虑了组间离差又考虑了组内离差。而主成分的第一个降维方向只考虑了总的偏差平方和。2) 在我们寻找偏最小二乘降维方向的时候，寻找的是响应变量与自变量相关性最大的方向[5]，而主成分降维法并没有考虑响应变量这一因素。因此在判别的准确率上了，偏最小二乘降维法得到的准确率显著地高于主成分降维法得到的数据。

另外，从几种常见的判别方法上来看。在测试集上，二次判别方法(QDA)较差，而且差生了过拟合现象。而支持向量机法(SVM)表现较好(见表 1)，在测试集上的准确最高达到了 83% (见表 1)。

从维数降低上来看，当 $m = 50$ 时的准确率总体上低于 $m = 20$ 的准确率(见表 1)，这说明，当我们选择更多的特征参与判别的时候，并不会一致地提高我们判别的准确性，这是因为，随着变量的增多，很有可能我们引入了更多与响应变量无关的噪音。进一步降低了我们分类的准确率。即：仅仅只有一少部分变量是真正与我们的响应变量有关的。

我们选取 $m = 20$，利用偏最小二乘(PLS)降维法，分别把维数降低到二维(图 1)和三维(图 2)，并在坐标系中将降维后的数据表示出来。两张图中，黑色的点代表未患病的样本点，红色的点代表患病的样本点。在图 1 中可以看到，两类样本可以用一条直线“粗略”地分开，即使在数据比较集中的左下角，两类数据也分别比较聚集，大概分别位于两个椭圆在之内。数据在第二个维度上的区分不是很明显，但是，在第一个偏最小二乘方向区分度较高，大概以取值−0.5 为分界线，取值比−0.5 大的大部分是健康的样本，反之则是患病的样本。说明了偏最小二乘降维法的第一个维度包含了大量的样本变异，能够比较精确地对样本进行分类，该维度对于样本的分类贡献最大。进一步可以知道，未患病的样本倾向于在第一个偏最小二乘方向上取得较大的数值(见图 1)。如图 2 所示，相对于二维图形，三维图形有着更好的分类效果，这说明增加了一个维度，相应地提高了我们的分类准确率。与二维图类似，我们可以找一张平面，把两类样本更为精确地区分开来。这也验证我们在表 1 中所显示的，线性判别法有着更高的分类正确率。

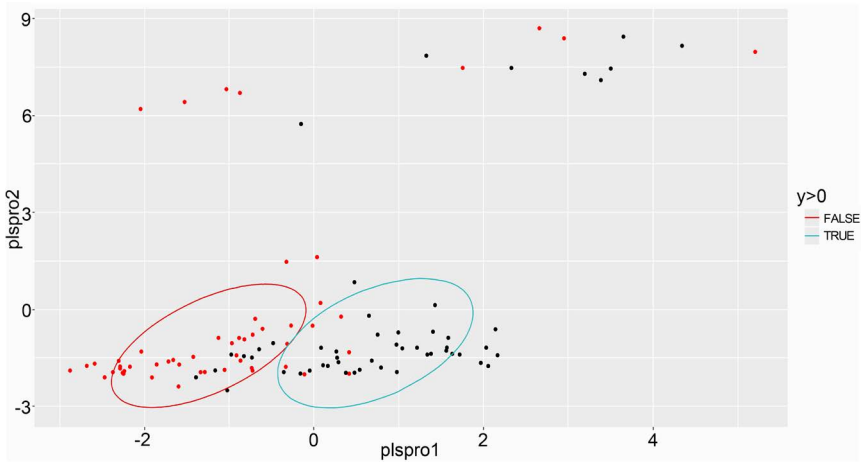

Figure 1. When $m = 20$, Illustration of dimension reduction for serum microRNAs data First 2 PLS gene components are employed. Black plots represent healthy samples, Red plots represent samples of illness
图 1. $m = 20$ 时，miRNAs 数据经过偏最小二乘法降维至二维后的图示

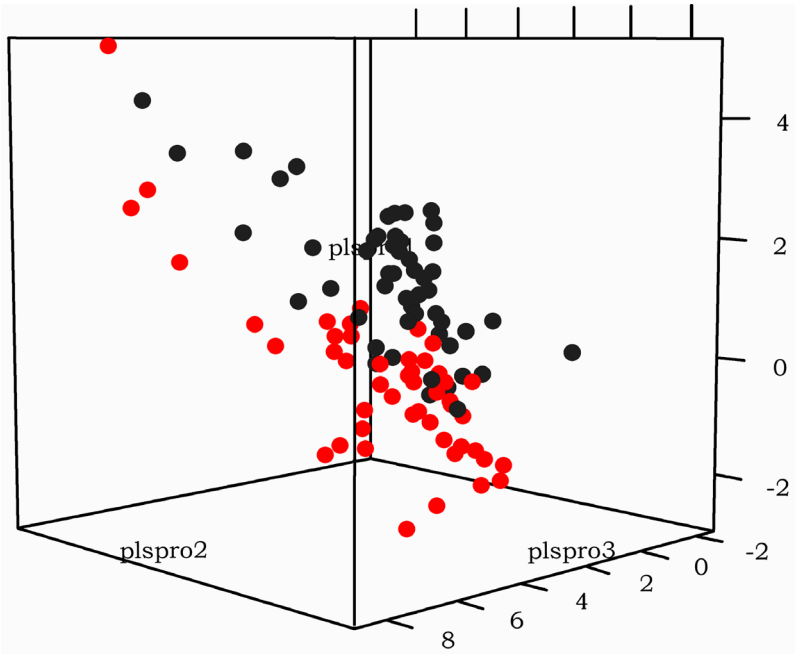

Figure 2. When $m = 20$, Illustration of dimension reduction for serum microRNAs data. First 3 PLS gene components are employed. Black plots and red plots represent healthy samples and samples of illness, respectively. The 3-dimensionaL PLS gene components plot illustrate the separability of the different classes
图 2. $m = 20$ 时，miRNAs 数据经过偏最小二乘法降维至三维后的图示

4. 讨论

利用与食管鳞癌相关血清 miRNAs 数据对疾病进行诊断在临床医学上具有重要是应用价值。本文所讨论的数据包括具有自变量数量多于样本数量、变量之间存在很强的多重共线性、分布特征不明确等问题，因此传统的分析方法得到的分类准确率较低。降维后的数据再进一步使用线性判别方法(LDA)或者支持向量机(SVM)等方法得到的分类准确率较高，也比较容易理解。但是它是利用原始变量的线性组合[8]作为新的自变量，虽然我们得到了较高的分类准确率，但是也存在变量解释上的困难。另外，在利用简单 t-统计量进行第一步降维的时候，是否筛选出来对于响应变量最有贡献的自变量这个问题上也存在着进一步优化的空间。能否可以通过惩罚型的偏最小二乘法选择最优的变量，在一步之中实现变量选择并且能完成变量的线性组合，进一步提高分类的准确率和变量的可解释性，并且能这种方法拓展到多元分类问题上，这也是作者当前所研究的问题。

基金项目

中国石油大学(北京)《概率论与数理统计》核心课程建设项目，项目编号 30YD1962。

参考文献

[1] Zheng,D.q. Ding, Y.j., Ma, Q. and Zhao, L. (2019) Identification of Serum MicroRNAs as Novel Biomarkers in Esophageal Squamous Cell Carcinoma Using Feature Selection Algorithms. *Frontiers in Oncology*, **8**, 674. https://doi.org/10.3389/fonc.2018.00674

[2] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[3] Liu, Y. and Wu, Y. (2007) Variable Selection via a Combination of the L_0 and L_1 Penalties. *Journal of Computational and Graphical Statistics* (Accepted for Publication).

[4] Barker, M. and Rayens, W. (2003) Partial Least Squares for Discrimination. *Journal of Chemometrics*, **17**, 166-173. https://doi.org/10.1002/cem.785

[5] Nguyen,D.V. and Rocke, D.M. (2002) Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data. Bioinformatics, **18**, 39-50. https://doi.org/10.1093/bioinformatics/18.1.39

[6] 高慧璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2005.

[7] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 统计学习导论[M]. 北京: 机械工业出版社, 2017.

[8] 王惠文. 偏最小二乘回归方法及其用[M]. 北京: 国防工业出版社, 2000.