

# 一种适于高维小样本数据的线性判别分析方法

成 忠, 诸爱士

(浙江科技学院 生物与化学工程学院, 杭州 310023)

**摘 要:** 针对高维小样本数据的类(模式)内离散度矩阵常为奇异,提出了一种改进的线性判别分析方法 ModLDA。它通过嵌入偏最小二乘算法,完成投影方向矢量的稳健估计,进而提取出若干个特征变量。而后基于特征变量张成的低维空间,构造样本类别的线性判别函数。在实证中,将 ModLDA 应用于药物光谱数据的化学模式识别,结果显示 ModLDA 方法判别能力明显优于其他方法。

**关键词:** 线性判别分析;特征提取;偏最小二乘;模式分类;药物光谱数据;模式识别

中图分类号: TP391.4;R911

文献标识码: A

文章编号: 1671-8798(2008)02-0098-04

## Modified linear discriminant analysis and its application in the small samples problem with high dimension

CHENG Zhong, ZHU Ai-shi

(School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China)

**Abstract:** In high dimensional and small sample size case, how to extract the optimal Fisher discriminant features efficiently remains unsolved. Now a novel classifier was constructed in this paper by combination of the non-linear iterative partial least squares algorithm with Fisher linear discriminant analysis (LDA). The resulting discriminate model based on this modified approach (ModLDA) was divided into two parts: the inner part that estimated the robust weight vector of canonical variates by linear partial least square algorithm and the outer part that built the LDA discriminate model in making use of the extracted canonical variates. The method utilized partial least squares regression as an engine for solving an eigenvector problem involving singular covariance matrices. Finally, application to a four-group problem with Raman spectroscopy data of the proposed ModLDA approach was presented with comparison to some other methods. The result demonstrates that limitations of LDA were overcome with partial least squares algorithm and then the classification performance of LDA was obviously improved.

**Key words:** linear discriminate analysis; feature selection; partial least squares; pattern classification; pharmaceutical tablet spectroscopy data; pattern recognition

---

收稿日期: 2008-04-10

基金项目: 浙江省自然科学基金资助项目(Y406053)

作者简介: 成 忠(1973—),男,江苏盐城人,副教授,博士,主要从事化学化工信息智能处理研究。

线性判别分析(linear discriminant analysis, LDA)作为一种便捷、有效的特征提取与数据分析方法,它基于 Fisher 准则,以样本的可分性为目标寻找一组线性变换,使样本类内离散度最小而类间离散度最大,因而在模式识别领域得到了广泛的应用<sup>[1-2]</sup>。然而,对高维小样本数据,样本类(模式)内离散度矩阵通常表现为奇异,应用 LDA 方法易产生过拟合现象。对此,目前常选用主成分分析(principal component analysis, PCA)集成 LDA 的两步策略<sup>[3-4]</sup>,即将 PCA 作为预处理步骤,以消除噪声及变量间的复共线性,保证投影后样本的类内离散度矩阵是非奇异的,并将得到若干个主成分(scores),随之组合为新的分类特征矢量。然而,PCA 提取的新分类特征矢量,未区别样本类内、类间信息,因此对于构建 LDA 判别模型通常并非为最优,而且需用于分类的成分仍然较多。

鉴此,本文将提出一种改进的线性判别分析(modified LDA, 简记为 ModLDA)方法,以实现高维小样本数据的降维、特征提取及模式分类。首先,它在施以 LDA 算法<sup>[5-6]</sup>进行高维数据的降维和特征提取时,通过嵌入偏最小二乘(partial least squares, PLS)算法<sup>[7]</sup>,实现投影方向矢量的稳健估计,进而提取出若干特征变量。这不仅可解决传统 LDA 方法在处理类内离散度矩阵为奇异时投影方向矢量计算的不稳健,又可以有效滤除噪音及变量间的复共线性。然后,在由特征变量张成的低维空间中,构造样本的线性判别函数,以完成模式分类器的设计。为考察 ModLDA 方法的有效性及性能,本文实施了药物(pharmaceutical tablet)拉曼光谱数据分类的应用试验研究。

## 1 ModLDA 方法设计

### 1.1 传统 LDA 方法及其缺点

对于  $n$  个样本的  $c$  类问题,假设  $i$  ( $i = 1, 2, \dots, c$ ) 类模式的先验概率为  $p(i)$ , 该类样本数为  $n_i$ 。样本属性矢量记为  $\mathbf{x} = (x_1, x_2, \dots, x_p)$ ,  $p$  为属性个数。LDA 作为一种线性降维技术,在实施 LDA 算法时,需先计算各模式类的平均分类属性矢量  $\bar{\mathbf{x}}_i$  和整个样本的平均分类属性矢量  $\bar{\mathbf{x}}$ , 并进而计算样本类内离散度矩阵  $\mathbf{S}_w$ 、类间离散度矩阵  $\mathbf{S}_b$ <sup>[5]</sup>, 如式(1)和式(2)所示。

$$\mathbf{S}_w = \frac{1}{n - c} \sum_{i=1}^c \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \quad (1)$$

$$\mathbf{S}_b = \frac{1}{c - 1} \sum_{i=1}^c n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \quad (2)$$

式(1)中:  $\mathbf{x}_{ij}$  为第  $i$  模式类样本中的第  $j$  个样本个体的分类属性矢量。然后, LDA 就是搜寻某一投影方向矩阵  $\mathbf{W} \in \mathbb{R}^{p \times r}$  ( $r$  为特征变量空间维数), 使得准则函数式(3)的取值最大。

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|} \quad (3)$$

在数学上, 容易证明使得上述准则函数最大化的  $\mathbf{W}$  的第  $k$  列向量  $\mathbf{w}_k \in \mathbb{R}^p$  ( $k = 1, 2, \dots, r$ ) 必须满足<sup>[5]</sup>:

$$\mathbf{S}_w \mathbf{w}_k = \lambda_k \mathbf{S}_b \mathbf{w}_k \quad (4)$$

式(4)即成为一个普通的本征值问题, 其中  $\lambda_k$  为最大本征值,  $\mathbf{w}_k$  为与其对应的本征矢量。由于  $\mathbf{S}_b$  的秩为  $c - 1$  或更低, 这样非零本征值的个数, 即对应于特征变量空间的维数, 将至多只有  $r = \min(p, c - 1)$  个。

然而, 对于高维小样本数据  $\mathbf{X} \in \mathbb{R}^{n \times p}$ , 常因其若干分类属性变量间存有较严重的复共线性, 由此计算的  $\mathbf{S}_w$  为奇异矩阵<sup>[2]</sup>, 再经式(4)求解的  $\mathbf{w}_k$  将不具稳定性, 进而由其计算特征变量  $t_k = \mathbf{X} \mathbf{w}_k$  也不具稳定性。

### 1.2 改进的 ModLDA 方法

现本研究选择 PLS 有偏估计方法<sup>[7]</sup>, 以实现投影方向矢量  $\mathbf{w}$  的稳健估计。将方程(4)等式左边  $\mathbf{S}_w \mathbf{w}$  的计算, 通过代入式(2)  $\mathbf{S}_b$  而重新拆分整理为  $(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \mathbf{w}$  等项乘积, 其中后一项实为常数, 其数值大小和  $\mathbf{S}_b$  计算式中的参数  $c, n_i$ 、方程(4)等式右边的  $\lambda_k$  等均不会影响  $\mathbf{w}_k$  的方向确定。这样, 对于  $c$  类问题, LDA 对投影方向矩阵  $\mathbf{W}$  的搜寻就可转化为多元线性回归方程系数矩阵  $\mathbf{B}$  的求解:

$$\mathbf{Y} = \mathbf{R} \mathbf{B} + \mathbf{F} \quad (5)$$

式(5)中:  $\mathbf{Y}$  的第  $i$  列向量为  $\mathbf{y}_i = \bar{\mathbf{x}}_i - \bar{\mathbf{x}}$ ;  $\mathbf{R} = \mathbf{S}_b$ ;  $\mathbf{B}$  为待估计系数阵;  $\mathbf{F}$  为残差矩阵。考虑到  $\mathbf{S}_w$  为奇异矩阵, 本研究对  $\mathbf{B}$  的稳健估计将交由多因变量的偏最小二乘 PLS2 算法<sup>[8]</sup>完成。

PLS2 算法, 常用非线性迭代偏最小二乘(nonlinear iterative partial least squares, NIPALS)方式<sup>[9]</sup>, 从  $\mathbf{R}$  和  $\mathbf{Y}$  中逐步成对地提取成分, 并进行线性回归。在由交叉验证法<sup>[10]</sup>确定最优成分数  $h$  后, 将得到  $\mathbf{R}$  的转换权阵  $\mathbf{V}$ 、载荷向量阵  $\mathbf{P}$ , 以及  $\mathbf{R}$  和  $\mathbf{Y}$  各成分对间的回归系数阵  $\mathbf{C}$ , 进而计算出式(5)的回归系数阵  $\mathbf{B} = (\mathbf{V}(\mathbf{P}^T \mathbf{V})^{-1} \mathbf{C}^T)_{R, Y}$ <sup>[11]</sup>。依次取出  $\mathbf{B}$  的每一列矢量  $\mathbf{b}_i$  ( $i = 1, 2, \dots, c$ ), 代入形同式(3)的目标

函数  $J(\mathbf{b}) = \frac{|\mathbf{b}_i \mathbf{S}_i \mathbf{b}_i|}{|\mathbf{b} \mathbf{S}_i \mathbf{b}|}$  进行计算, 并将它们按函数值的大小排序, 再从排序后的  $\mathbf{B}^*$  中剔除最后一列, 即第  $c$  列矢量, 便将得到稳健的  $\mathbf{W}$ 。

基于 PLS2 算法提取的  $\mathbf{W}$  计算特征变量:  $\mathbf{T} = \mathbf{XW}$ , 进而用于构造 LDA 的判别函数。首先, 它计算样本类内离散度矩阵  $\mathbf{S}_{w,T}$ , 然后构造第  $i$  类别 LDA 的判别函数(类模型)<sup>[11]</sup>:

$$L(\mathbf{t}) = \log(p(i)) - \frac{1}{2}(\mathbf{t} - \bar{\mathbf{t}}) \mathbf{S}_{w,T}^{-1}(\mathbf{t} - \bar{\mathbf{t}}) + \log|\mathbf{S}_{w,T}| \tag{6}$$

式(6)中:  $\mathbf{t}$  为特征变量空间中待判别的分类属性矢量;  $\bar{\mathbf{t}}$  则为特征变量空间中第  $i$  类模式的平均分类属性矢量。对于  $c$  类问题, 分类器则视为一个计算形如式(6)的  $c$  个判别函数, 并选取与最大判别值对应的类别的机器。

现基于训练样本矩阵  $\mathbf{X}$ , 将改进的 ModLDA 算法步骤归结如下:

- 1) 计算各模式类的平均分类属性矢量  $\bar{\mathbf{x}}_i$  和整个样本的平均分类属性矢量  $\bar{\mathbf{x}}$ ;
- 2) 构造  $\mathbf{Y}$  矩阵, 其第  $i$  列 ( $i = 1, 2, \dots, c$ ) 向量  $\mathbf{y}_i = \bar{\mathbf{x}}_i - \bar{\mathbf{x}}$ ;
- 3) 通过式(1) 计算类内离散度矩阵  $\mathbf{S}_w$ ;
- 4) 组合新样本集  $\{\mathbf{S}_w, \mathbf{Y}\}$ , 对其实施 PLS2 算法以解决投影方向矩阵  $\mathbf{W}$ ;
- 5) 计算特征变量  $\mathbf{T} = \mathbf{XW}$ ;
- 6) 对一样本个体  $\mathbf{x}$  为  $\mathbf{R}$ , 从  $\mathbf{W}$  中依次取出第  $k$  列  $\mathbf{w}_k$ ,  $k = 1, 2, \dots, r$ , 以计算  $\mathbf{t}_k = \mathbf{xw}_k$  后组合为  $\mathbf{t}$ ;
- 7) 将  $\mathbf{t}$  代入式(6) 各类别的 LDA 判别函数 ( $c$  个), 判定最大判别值对应的类别。

## 2 ModLDA 应用与结果分析

### 2.1 试验数据说明

药物 (pharmaceutical tablet) 数据取自文献 [12], 分类属性取其拉曼光谱在 [200, 3 600 nm] 波长区间、间隔为 1 nm 的若干波长处的拉曼强度, 维数  $p = 3\,401$ , 如图 1 所示。其自变量相关矩阵  $\mathbf{X}^T \mathbf{X}$  的 3 401 个特征根中近 3 281 个因小于  $9.38 \times 10^{-9}$  而非常接近于 0, 表明自变量间存在严重的复共线性。类别属性则为药物剂量 (dosages) 不同包装规格: 5, 10, 15, 20 mg 共 4 类, 并分别以数字 1, 2, 3, 4 标识。而这 4 模式类个体数目分别为 30, 27, 33, 30, 即样本容量  $n = 120$ 。另外, 在建模前, 对各分类

属性变量作了中心化预处理。

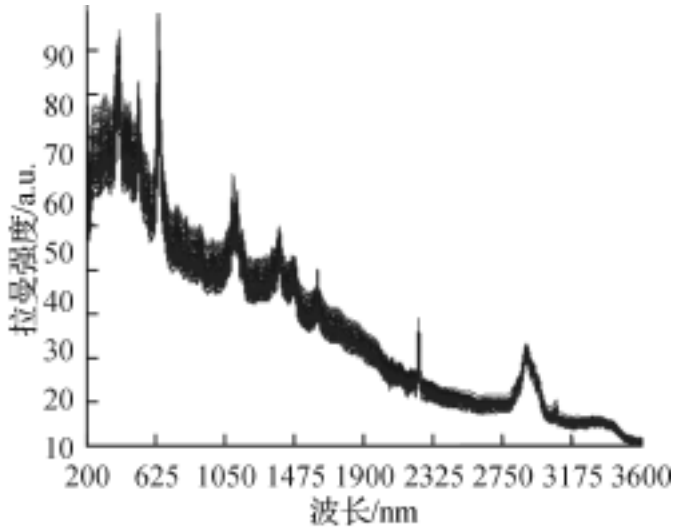


图 1 药物样本的拉曼光谱

Fig. 1 Raman spectroscopy of pharmaceutical tablet

### 2.2 试验结果与分析

2.2.1 分类器分类性能比较与分析 为避免偶然性, 试验选择以“five-fold 交叉验证”方法, 从样本中依次取出若干个, 即  $n/5$  个组成预测样本, 剩余  $4n/5$  个用于训练, 共 5 轮。分类器的性能, 则以对训练样本的分类正确率称为自检正确率 Crt-S (correctness rate of self-checking) 和对预测样本的分类正确率称为预测正确率 Crt-P (correctness rate of predicting) 来表征<sup>[13]</sup>。同时选用了 LDA 和 PCA-LDA 方法<sup>[14]</sup> 建模, 以作比较。3 种分类器的性能比较结果列于表 1, 它们均为 5 轮试验的平均值。另外, ModLDA 方法中, PLS 最优成分数  $h = 10$ , 特征变量个数  $r = 3$ 。为便于比较, PCA 和 LDA 的特征变量数也取值为 3。

表 1 3 个分类器的模型识别性能比较

Table 1 Comparison on recognition performance of the three classifiers

分类器	Crt-S/ %	Crt-P/ %
LDA	97.92	85.83
PCA-LDA	62.50	53.33
ModLDA	91.67	90.83

由表 1 可见, 对于药物剂量的分类结果, ModLDA 分类器的 Crt-P 为最高, 这表明 ModLDA 通过嵌入的 PLS2 算法实现了  $\mathbf{W}$  的稳健估计, 从而提取了分类能力更优的特征变量。LDA 分类器的 Crt-S 不仅明显高于另两种分类器, 也较其自身的 Crt-P 为高, 这可能为训练样本过拟合的结果。而在取相同的  $r = 3$ , PCA-LDA 的 Crt-P 和 Crt-S 均为最低, 这是由于在成分提取的步骤中, 未能区别样本先验的类别信息, 因此获取的 PCA 成分, 即新的分类属

性对于构建 LDA 判别模型并非为最优。若将 PCA 成分数升至其最优成分数 13, 则其 Crt-P 将增至最大 80.83%, 而 Crt-S 为 93.96%, 模型的预测判别能力仍不及 LDA 和 ModLDA 方法, 且模型结构因特征维数的增大而变得复杂。

### 2.2.2 特征权矢量 $\mathbf{w}$ 的比较与分析

图 2 为 3 个

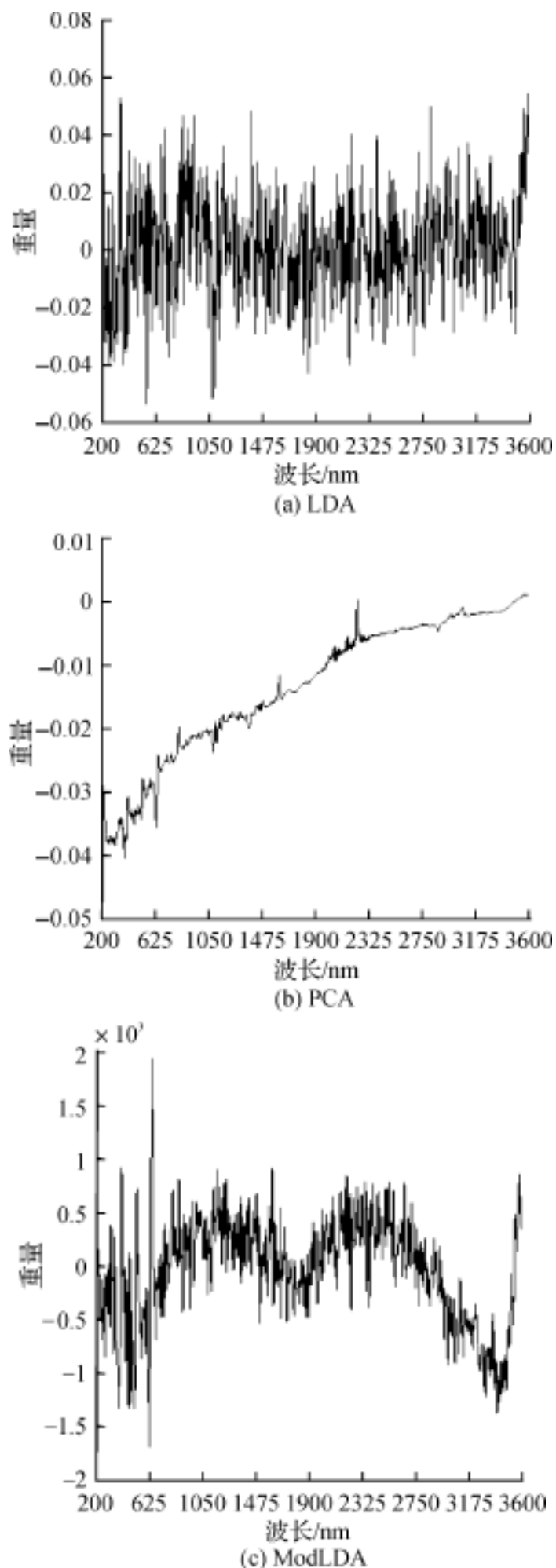


图 2 第一个投影方向矢量

Fig 2 The first mapping direction vector for (a) LDA (b) PCA (c) Mod LDA

分类器, 在断点式交叉验证的第一轮中, 分别经 LDA、PCA 和 ModLDA 降维算法所获得的第一个特征权矢量  $\mathbf{w}$ 。在这 3 种降维算法中, 后继计算  $\mathbf{t}$  的算式可以统一为  $\mathbf{t} = \mathbf{X}\mathbf{w}$ , 由此可以将  $\mathbf{w}$  的物理意义理解为, 它的分量  $w_s$  ( $s = 1, 2, \dots, p$ ) 显示变量  $x_s$  对成分  $\mathbf{t}$  的贡献, 其值越大, 表明该变量的信息就越重要, 这也正是上述方法常被用作谱图波长选择的缘由。

从图 2(b) PCA 和图 2(c) ModLDA 产生的  $\mathbf{w}$  来看, 对类别判定有效的信息波长点被划分为几个小的区段, 这几个区段的加权系数在数值上存在着明显的差异, 且后者 ModLDA 的区分能力更为突出。而在图 2(a) 中, LDA 产生的  $\mathbf{w}$  在全波长范围内变化剧烈, 对波长的选择性不明显, 这正是  $\mathbf{w}$  估计不稳定所致。由此可以得到, ModLDA 方法提高了类别对波长的选择能力, 从而提高了判别模型分类精度和稳定性。

### 2.2.3 ModLDA 分类器参数的选择与确定

影响 ModLDA 模型分类正确率的主要因素有: 特征变量个数  $r$ 、PLS 成分数  $h$ , 以及各类别的先验概率  $p(i)$  等。由于本文的试验实例中没有提供各类别的先验概率, 故一般简单假设每一类的先验概率相等, 即四分类问题取值为  $p(i) = 1/4$  ( $i = 1, 2, 3, 4$ )。特征变量个数  $r$ , 对高维样本  $p \gg m$  来说, 由 1.1 节分析可知, 一般取值为  $r = c - 1$ , 如本文的四分类问题  $r = 3$ 。PLS 成分数  $h$ , 往往由交叉验证方式下, 分类器对预测样本的最低分类错误个数优化确定。

## 3 结 语

本研究改进的 ModLDA 方法, 通过在 LDA 方法中集成 PLS 算法, 完成初始高维特征空间的稳健降维, 同时有效消除采样中的光谱数据噪声及自变量间的复共线性, 并得到一组特征变量。然后, 基于特征变量张成的低维特征空间, 通过 Fish 线性判别分析, 完成模式分类器的构建, 并进而用于样本类别的判定。将 ModLDA 方法应用于药物剂量类别属性与其拉曼光谱数据的化学模式识别, 降维效果显著, 特征变量估计稳健, 分类器误判率低。

(下转第 113 页)