多态偏最小二乘法模型①

徐继刚[1], 冯新泸[1], 管 亮[1], 王 帅[1], 杨 岚[2]

[1](后勤工程学院 军事油料应用与管理工程系，重庆 401311)
[2](北空后军需物资油料处，北京 101400)

摘 要：为了更合理的确定偏最小二乘法的主成分数，提出了一种多态偏最小二乘法的建模方式。介绍了建模和预测具体实现过程。给出了预测时样品相似度计算的两种方式：直接距离法和性质得分距离法。以玉米样品近红外光谱数据为例，分别采用多态偏最小二乘法与传统偏最小二乘法建模对蛋白质指标进行了检测。结果表明：多态偏最小二乘法预测结果优于传统偏最小二乘法预测结果,有更强的适应性和兼容性。

关键词：偏最小二乘法；模型；主成分数；相似性度量

Polymorphic Partial Least Squares Model

XU Ji-Gang[1], FENG Xin-Lu[1], GUAN Liang[1], WANG Shuai[1], YANG Lan[2]

[1](Dept. of Oil Application & Management Ngnineering, LEU, Chongqing 401311, China)
[2](Dept of Beijing Military Area Air Force, Beijing 116000, China)

Abstract：In order to determine the principal component number of Partial least squares(PLS), we propose a polymorphic PLS modeling approach. Describes the specific process of building model and predictive, two methods of sample similarity calculation is gived, while we predict samples: direct distance and property score distance. The article takes the near infrared spectroscopy(NIR) data of corn Samples as example, and detects the protein by polymorphic PLS model and traditional PLS model. The results indicated that: prediction results of the polymorphic PLS is better than prediction results of traditional PLS, it has greater flexibility and compatibility.

Key words：partial least squares; model; principal component number; similarity measure

传统偏最小二乘法因能同时将因变量矩阵和自变量矩阵用主成分表示，充分表现并利用因变量矩阵和自变量矩阵的信息，因此在数据建模中被广泛应用。然而很多研究者也发现偏最小二乘法还存在一些缺陷，比如主成分选取非最优化和拟合过度等问题[1-3]。人们在具体应用过程中对传统的偏最小二乘法进行一些改进，助推 PLS[3]，间隔 PLS[4-7]，核 PLS[8,9]，多模型共识的 PLS[10]，多项式 PLS[11]和局部加权 PLS[1,12]等建模方式被提出和应用，在解决线性和非线性问题上比传统偏最小二乘法效果更好。本文主要从偏最小二乘法主成分数确定的角度出发，对传统偏最小二乘法进行改进，传统偏最小二乘法是找到使得训练集所有样品预测残差平方和最小的主成分数为模型的主成分数，采用留一法全交互验证实现，这样的主成分数包含了训练集样品的统计信息，模型的兼容性较强，但是针对某个未知预测对象来说，这种方式得到的模型未必能得到最好的预测效果[1,2]。由于不同样品的特征不同，对不同的样品，最优主成分数是不同的，因此我们对多主成分数偏最小二乘法[5]进行了研究：根据样品的特征不同为训练集中的每个样品确定对应的最优主成分数，采用这样的主成分数构建的模型相应的就会对和样品相似度高的未知样品有好的预测结果。这样的模型针对性较强，当未知样品与训练集中样品相似性都较差时，多主成分数偏最小二乘法就会失去优势，这时采用传统偏最小二乘法会得到好的预测效果。为了得到综合性强的预测效果，将两种主成

① 收稿时间:2011-09-29;收到修改稿时间:2011-12-19

分数确定方式结合起来，构建一种新的偏最小二乘法模型，因为在形式上与程序设计中的多态比较像，故称之为多态偏最小二乘模型。

1 模型的构建与预测方法

1.1 传统偏最小二乘法模型和多主成分数偏最小二乘法模型的构建

1.1.1 传统偏最小二乘法的基本原理

为了研究因变量与自变量的统计关系，观测了 n 个样本点，由此构成了自变量与因变量的数据表 X 和 Y，X 为自变量观测值矩阵，Y 为因变量观测矩阵值。偏最小二乘回归分别在 X 与 Y 中提取出 t 向量和 u 向量，要求：(1) t 和 u 应尽可能大地携带它们各自数据表中的变异信息;(2)t 和 u 的相关程度能够达到最大。在第一个成分被提取后，偏最小二乘回归分别实施 X 对 t 的回归以及 Y 对 u 的回归。如果回归方程已经达到满意的精度，则算法终止；否则，将利用 X 被 t 解释后的残余信息以及 Y 被 t 解释后的残余信息进行第二轮的成分提取。如此往复，直到能达到一个较满意的精度为止。

1.1.2 传统偏最小二乘法主成分数的确定及构建

在对自变量 X 和因变量 Y 进行 PLS 建模过程中，先对因变量 X 矩阵求解特征值和特征矢量，化学上，特征矢量我们称为主成分，按照特征值的大从大到小，特征矢量分别为第 1 主成分，第 2 主成分，…第 n 主成分。每个主成分对预测结果的贡献不同，在具体建模过程中要确定具体采用前几个主成分进行建模，这个过程就是确定主成分数。具体的确定主成分数的方法如下：

对于因变量为 Yn，自变量为 Xn×m 的训练集，样品数为 n，数据点数为 m。

（1）在训练集中取出第 k（k 的初始值设为 1）个样品数据为待测样本 yk,x1,k 这时剩余样品构成训练集矩阵为：Yn-1，X（n-1）×（m-1）,设主成分数初值为 P=1；

（2）用 Yn-1，X（n-1）×（m-1）训练集和前 p 个主成分建立回归模型；

（3）用第（2）步建立的数学模型去预测第（1）步取出的样本，通过预测值 y 来计算残差平方 PRES：$PRES(1)=(y_k-y)^2$。

（4）循环迭代，k=k+1，返回第(1)步，把每一个训练集样本轮流进行预测，直到 k=n，把训练集每一个样本的残差平方和累加，得到第 p 个主成分的预测残差平方和 $PRESS(p)=\sum_{i=1}^{n}PRES(1)$；

（5）假如 p<m,p=p+1,重复第（1）步到第（4）步。

这样便可以求得前 1 到 m 个主成分时的预测残差平方和。$BestNum=\arg(\min(PRESS(m)))$。即 PRESS 达到最小时的主成分数确定为该训练集建模的最优主成分数 BestNum，采用这时的回归参数为建模参数。

1.1.3 多主成分数最小二乘法构建

传统偏最小二乘法建立的模型，只为模型确定一个主成分数，一组回归参数。多主成分数偏最小二乘法，为每个训练集的每样品确定一个主成分数，得到对应的一组回归系数。具体流程图，如图 1。

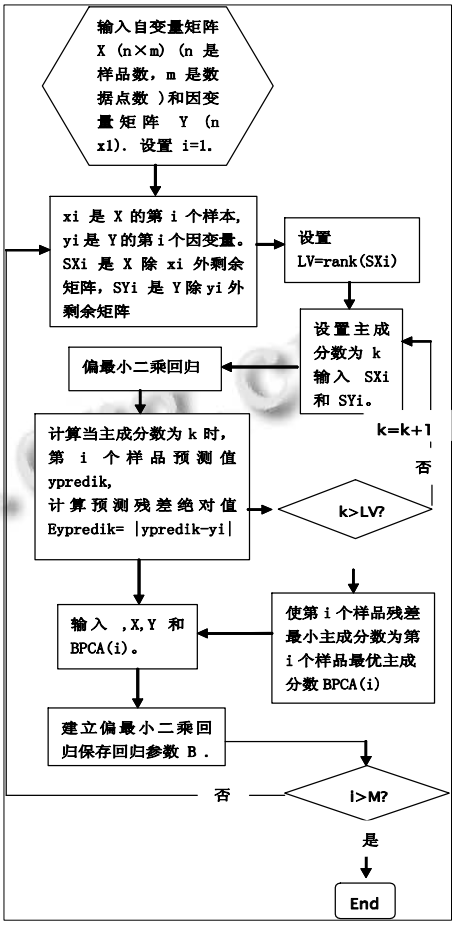

图 1 多主成分数偏最小二乘法校正模型算法流程图

1.2 相似性度量的计算

多态偏最小二乘法就是在未知样品预测时，在传

统偏最小二乘法和多主成分数偏最小二乘法中选择一个更好的预测模型。选择的依据就是未知样品与训练集中样品的最小距离大于训练集样品间的最大距离时，采用传统偏最小二乘法回归模型计算，否则，采用多主成分数偏最小二乘法计算，通过相似度找到训练集样品中距离最近的样品，用该样品的最佳主成分数对应的回归参数进行预测。

样本相似性的度量一直是很难精确定义的问题，因此相似性度量对于有效地利用模型至关重要。在样品模式空间对相似度的度量引用的几何中的距离[13]。常用的距离有马氏距离，欧式距离，汉明距离，曼哈顿距离，夹角的余弦，相关系数等。在实际应用中，有时相似度和性质指标是相关的，对于不同的指标相似度是不同的。因此在相似度计算时采用两种方式，第一种样品间直接距离法，第二种是性质得分距离法，具体算法是预测样本与传统偏最小二乘法最佳主成分数对应的载荷矩阵相乘得到预测样本的得分向量，再通过具体的距离计算预测样本得分向量与最佳主成分数对应的训练集各样本的得分向量间的距离。模型应用时可以根据具体的情况选择不同的距离计算方法。

1.3 多态偏最小二乘法模型构建及预测过程

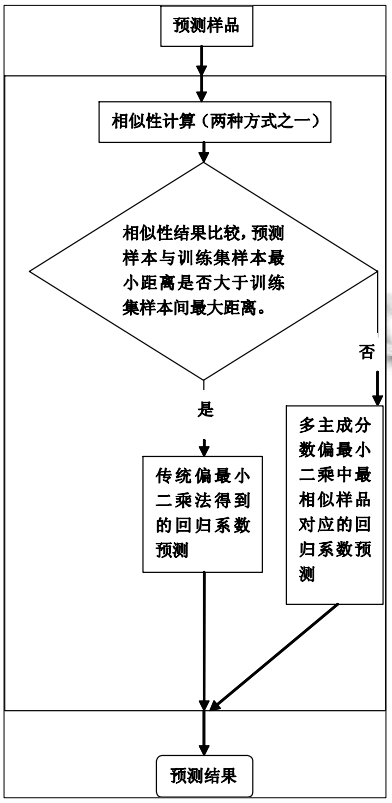

图 2　多态式偏最小二乘法预测过程

因为在样本预测时要同训练集的样品进行对比，所以多态偏最小二乘模型构建时要保存许多训练集样本信息和中间过程的一些结构化的参数信息。具体包括：训练集样本的自变量数据，因变数据和各种相似度算法和参数下的样本间距离最大值，传统偏最小二乘法的最优主成分数载荷矩阵，得分矩阵，回归系数和各种算法和参数下得分矩阵距离最大值，多主成分数偏最小二乘法得到的每个样品最优主成分数及对应的回归系数，相似性度量的计算函数（直接距离法和性质得分距离法）及距离算法参数（马氏距离，欧式距离，曼哈顿等）。具体模型的预测过程如图 2。

2　在近红外光谱测定玉米的蛋白质含量中的应用

为了对比传统偏最小二乘法与多态式最小二乘法的预测结果，对一组玉米的近红外光谱进行了测试。实验数据来源：http://software.eigenvecto1.com/Data/Corn/index.html. 以原数据集中 m5 仪器上测量得到的一组近红外光谱数据用于测试，数据内容为：80 个玉米样品的近红外谱图，波长范围是 1100 到 2498nm,间隔 2nm。原始数据的近红外图，如图 3。

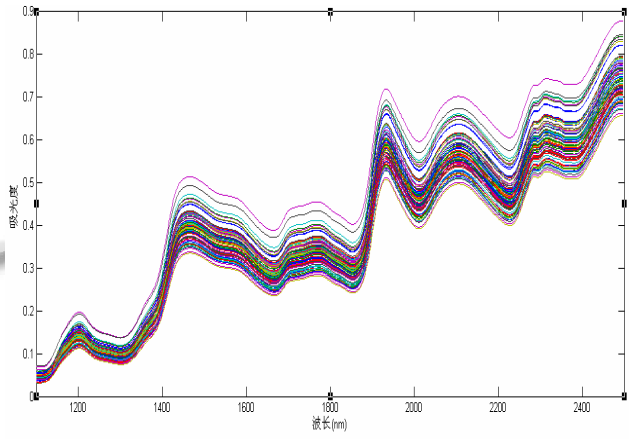

图 3　玉米近红外光谱原始数据谱图

分别采用传统偏最小二乘法和多态偏最小二乘法对样品的蛋白质指标进行建模预测。以 60 个样品为训练集，20 个样品为预测集。程序实现采用 MATLAB 2008 完成。传统偏最小二乘法建模的最优主成分数为 37，相似性量度采用的是性质得分法，选用的距离算法是曼哈顿距离。预测结果如表 1，从表 1 中可以看出除样品 7 外，多态偏最小二乘法的误差均小于或等

于传统偏最小二乘法的预测误差。多态偏最小二乘法预测时，当主成分数取 37 时，为预测样品在训练集样品中没找到相似性满意的样品（即预测样品与训练集样品间的最小距离大于训练集样品间的最大距离）。从预测结果来看，多态偏最小二乘法对预测结果有显著的的提高。性质指标得分和曼哈顿距离作为玉米的蛋白质指标预测的相似性度量方法是适用的。

表 1 两种建模方法的蛋白质指标预测结果

预测样品序号	实测值	传统 PLS 预测值	误差绝对值	多态 PLS 预测值	误差绝对值	最优主成分数
1	8.986	8.9433	0.0427	8.9433	0.0427	37
2	9.313	9.1597	0.1532	9.2155	0.0975	11
3	8.905	8.7905	0.1144	8.7906	0.1144	37
4	8.338	8.2770	0.0610	8.2770	0.0610	37
5	8.571	8.5746	0.0036	8.5746	0.0036	37
6	9.354	9.2333	0.1206	9.2334	0.1206	37
7	9.021	9.0791	0.0581	9.0996	0.0786	19
8	9.382	9.3425	0.0395	9.3425	0.0395	37
9	7.654	7.6740	0.0201	7.6677	0.0137	19
10	7.908	7.8698	0.0382	7.8698	0.0382	37
11	8.613	8.8466	0.2336	8.8393	0.2263	18
12	8.838	9.0930	0.2551	8.7705	0.0675	10
13	7.873	7.7708	0.1022	7.7708	0.1022	37
14	8.288	8.1647	0.1233	8.1647	0.1233	37
15	8.649	8.4173	0.2317	8.4173	0.2317	37
16	7.876	7.7763	0.0997	7.8265	0.0495	19
17	8.586	8.3286	0.2574	8.3286	0.2574	37
18	8.030	7.9939	0.0360	8.0420	0.0120	10
19	8.132	8.0137	0.1182	8.0434	0.0886	19
20	8.428	8.1369	0.2911	8.1464	0.2816	19

3 结语

给出了多态偏最小二乘法的建模与预测的具体实现方法，将多主成分数和样品的预测效果紧密的联系在一起，提高了模型的针对性和适应性，并保持原有的兼容性。以玉米的近红外光谱数据为例对蛋白质指标进行了预测，对多态偏最小二乘法的预测能力进行了验证，多态偏最小二乘法比传统偏最小二乘法有更好的预测能力。

参考文献

1 李军会,秦西云,张文娟.局部偏最小二乘回归建模参数对近红外检测结果的影响研究.光谱学与光谱分析,2007,27(2):262–264.

2 杨岚,冯新泸.动态优化偏最小二乘模型的建立与应用后勤工程学院学报,2008,24(2):75–77.

3 段宏博.助推偏最小二乘法(BPLS)及其应用.数学理论与应用,2009,29(4):118–121.

4 屠振华,籍保平,孟超英,朱大洲,史波林,庆兆坤.基于遗传算法和间隔偏最小二乘的苹果硬度特征波长分析研究.光谱学与光谱分析,2009,29(10):2760–2764.

5 石吉勇,邹小波,赵杰文,殷晓平.基于小波滤噪和 iPLS 的草莓近红外光谱糖度检测模型.安徽农业科学,5752–5754.

6 李艳肖,邹小波,董英.用遗传区间偏最小二乘法建立苹果糖度近红外光谱模型.光谱学与光谱分析,2007,27(10):2001–2004.

7 李振庆,黄梅珍,倪一,丁海峰,汤洁蔚,窦晓鸣. 改进偏最小二乘法在近红外牛奶成分测量中的应用.光学技术,2009,35(1):70–73.

8 吴炜,杨晓敏,余艳梅,石一兴,何小海.核偏最小二乘算法的图像超分辨率算法.电子科技大学学报, 2011,40(1):105–110.

9 朱红求,阳春华,桂卫华.基于 KPLS 和 LS-SVM 的过程参数预测及其应用.控制工程, 2010,17(2):216–218.

10 李艳坤,邵学广,蔡文生. 基于多模型共识的偏最小二乘法用于近红外光谱定量分析.高等学校化学学报,2007,28(2):246–249.

11 张琳,张黎明,李燕,王晓斐,胡兰萍,王俊德.多项式偏最小二乘法对非线性体系红外谱图的分析.光谱学与光谱分析 2006,26(4):620–623.

12 张莹,王耀南.基于局部加权偏最小二乘法的冷凝器污垢预测.仪器仪表学报, 2010,31(2):299–304.

13 梁逸曾,俞汝勤。分析化学手册第十分册化学计量学(第二版).北京:化学工业出版社,2000:328–330.