聂斌,杜玉文,杜建强,等. 融入距离方差和距离相关系数的偏最小二乘回归方法[J]. 江西师范大学学报(自然科学版),2023,47(1):61-68.
NIE Bin,DU Yuwen,DU Jianqiang,et al. The regression Method of PLS fusing distance variance and distance correlation coefficient[J]. Journal of Jiangxi Normal University(Natural Science),2023,47(1):61-68.

文章编号:1000-5862(2023)01-0061-08

融入距离方差和距离相关系数的偏最小二乘回归方法

聂 斌,杜玉文,杜建强*,张玉超,郑学鹏,靳海科
(江西中医药大学计算机学院,江西 南昌 330004)

摘要: 偏最小二乘法(partial least square,PLS)在内部采用 Pearson 系数度量自变量和因变量之间的相关性时提取出的成分不能确保解释性最强,并且 PLS 在将提取的成分进行线性回归时也无法真实反映变量间的函数关系. 针对这些问题,该文提出了融入距离方差和距离相关系数的偏最小二乘回归方法(DVDC-CPLS). DVDCCPLS 基于距离方差和距离相关系数提取距离成分,再将距离成分进行拟线性回归得到距离回归方程,通过模型求解方法将距离回归方程转换为原始数据的表达,最终得到结构简洁、精度较高的回归模型. 该文分别采用麻杏石甘汤数据和 UCI 数据集测试 DVDCCPLS 的性能,并与其他 5 种经典的回归算法对比,结果表明:DVDCCPLS 具有较好的回归效果和回归性能.

关键词: 偏最小二乘;距离方差;距离相关系数;回归方程;拟线性
中图分类号: TP 311 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2023.01.08

0 引言

偏最小二乘法(partial least squares,PLS)[1]是集主成分分析、典型相关分析和多元线性回归分析于一体的多元线性统计分析方法,它通过多个自变量对多个因变量之间的关系进行建模. PLS 首先在预测矩阵 X 中找到方向向量 w,用于解释响应矩阵 Y 的最大方差,然后通过获得的系数 w 将 X 和 Y 投影到新空间来建立线性回归模型. 当预测矩阵的变量比观测值多且观测值 X、Y 之间存在多重共线性时,PLS 回归尤其适用[2]. 因此,PLS 已被广泛应用于化学、经济学、药学、计算机视觉和模式识别等诸多领域[3].

应用 PLS 回归分析建立的模型是一种多元线性回归模型,要求自变量与因变量间有显著的线性关系,然而在实际应用中,变量之间往往不是严格的线性关系,而是复杂的非线性关系,且在实际数据中需要解决的往往不是单一的线性问题,而是线性和非线性均存在的复杂问题. 因此,若还是运用 PLS 对非线性问题或复杂问题进行建模,则将得到较低的回归精度,从而限制了 PLS 在非线性领域问题上的应用.

针对 PLS 对非线性数据不敏感的问题,有许多非线性 PLS 的改进方法被广泛提出. Liu Hongbin 等[4]使用高斯过程回归(GPR)在 PLS 中建立每对潜变量之间的非线性回归,提出一种动态 GPR-PLS 模型来提高估计能力. 尚栋等[5]构造不同影响变量的 3 次多项式将其加入 PLS 的建模中以近似校正非

收稿日期:2022-11-02
基金项目:国家自然科学基金(82260849,62141202),民族药资源数据库与信息网络化共享平台构建(2019YFC1712301)和江西中医药大学校级科技创新团队发展计划(CXTD22015)资助项目.
作者简介:聂 斌(1972—),男,江西峡江人,教授,博士研究生,主要从事数据挖掘、中医药信息学和中药学的研究. E-mail:ncunb@163.com
通信作者:杜建强(1968—),男,江西南昌人,教授,博士,博士导师,主要从事中医药信息与数据挖掘的研究. E-mail:jianqiang_du@163.com

线性因素带来的影响,提出一种基于循环变量筛选的非线性 PLS 方法. Ma Hao 等[6] 将一种由自回归外生模型和径向基函数神经网络组成的新型级联结构作为传统 PLS 方法的内部模型,使新建立的 PLS 方法具有动态和非线性的特性. 贾润达等[7] 和 Jiao Jianfang 等[8] 运用核函数将低维空间的非线性关系映射到高维空间的线性关系,使得 PLS 方法适用于非线性结构. Zhu Bao 等[9] 和 Wang Yanxia 等[10] 将人工神经网络与 PLS 集成在一起,实现一种新的鲁棒非线性 PLS 处理非线性和共线性数据的方法. 鲁庆华等[11] 提出了一种基于偏最小二乘回归和多项式回归相结合的多元非线性回归分析方法,该方法通过 PLS 提取的主成分进行多项式回归,实现了 PLS 的非线性回归. A. Merino 等[12] 通过增广输入矩阵包含基于知识变量的非线性变换特点,将增广输入矩阵进行 PLS 回归,提出了一种基于知识的递归 PLS 非线性回归方法. 李雄威等[13] 通过相关性分析选取与因变量相关的自变量,然后构建能够表征非线性关系的输入变量,从而得到针对具体问题的非线性表达式,再利用 PLS 方法建立模型得到表达式的系数,提出一种非线性 PLS 模型,但该方法构建的非线性表达式较简单. F. B. Lavoie 等[14] 通过约束谢文龙[15] 提出的 3 次样条分段迭代模型建立非线性关系,再通过 U. Indahl[16] 和 F. B. Lavoie 等[17] 提出的动力修正 PLS 迭代计算更新权重,提出了一种新的鲁棒非线性回归方法. Peng Shan 等[18] 提出了一种基于分段线性内部关系的非线性偏最小二乘(PLS) 切片变换(SLT) 模型,该方法使用基于分段线性映射函数的 SLT 构建输入和输出分数向量之间的非线性内部关系.

以上研究均在一定程度上缓解了 PLS 对非线性数据表现不佳的问题,大多数的非线性 PLS 改进模型均从外部学习数据的非线性特征作为 PLS 输入得到非线性模型,或者将原始特征进行变换,构建特征的拟线性表达式,再构建 PLS 的非线性回归模型. 然而,这些方法并未从内部改变 PLS 的线性框架,如提取的成分为原始数据的线性组合,内部采用 Pearson 系数衡量自变量和因变量的相关性,以及对提取的成分进行线性回归. 针对上述问题,本文基于 PLS 的思想提出一种新的回归方法,该方法通过距离方差反映原数据信息,内部采用距离相关系数度量自变量与因变量的相关性,在多元线性回归分析中采用拟线性回归[19] 方法.

1 融入距离方差和距离相关系数的偏最小二乘回归模型

1.1 偏最小二乘法(partial least square,PLS)

偏最小二乘法[20] 是集主成分分析、典型相关性分析和多元线性回归分析于一体的多元线性统计分析方法,可以解决多自变量与多因变量之间的建模问题,对于变量间多重相关性和小样本容量问题尤其适用.

假设一组自变量 $X=(x_1,x_2,\cdots,x_m)$ 和因变量 $Y=(y_1,y_2,\cdots,y_n)$,PLS 首先分别提取自变量第 1 主成分 t_1 和因变量第 1 主成分 u_1,提取的条件为方差 $D(t_1)\to\max$,$D(u_1)\to\max$,同时相关系数 $r(t_1,u_1)\to\max$;然后将提取的成分 t_1 和 u_1 进行线性回归,得到残差矩阵,判断是否满足精度要求,若精度满足,则算法停止,否则,利用残差矩阵提取下一个主成分,不断迭代直到满足精度要求;最后将提取得到的 h 个主成分 $t_1,t_2,\cdots,t_h(h<m)$ 进行线性回归,构建每个因变量 y_n 对 $t_1,t_2,\cdots,t_h$ 的回归方程,由于提取的主成分均为原始变量的线性组合,因此最终可表达为 y_n 关于原始自变量 $X=(x_1,x_2,\cdots,x_m)$ 的回归方程.

1.2 融入距离方差和距离相关系数的偏最小二乘回归算法

DVDCCPLS 算法首先将数据进行预处理,再通过距离方差和距离相关系数最大化提取距离成分,最后通过拟线性回归方法对距离成分进行回归,具体的算法构造过程如图 1 所示.

首先,将原始数据进行标准化得到 E、F,再将标准化后的矩阵计算样本间的欧氏距离得到实对称距离矩阵,再将距离矩阵进行中心化得到矩阵 A、B.

其次,根据距离方差和距离相关系数最大化的要求,分别求取方向向量 w_i 和 v_i 得到自变量距离成分 t_i 和因变量距离成分 u_i,图 1 第 2 部分 $D(Aw_i)$ 和 $D(Bv_i)$ 表示距离方差,$r(\cdot,\cdot)$ 表示距离相关系数;将得到的距离成分进行回归并计算距离残差和交叉有效性系数 Q_h^2,利用 Q_h^2 判断应提取的距离成分个数,其中 $S_{\text{press},h}$ 和 $S_{\text{ss},h-1}$ 分别表示预测误差平方和与误差平方和;最终判断模型是否满足精度条件,若是则输出提取的距离成分,若否则利用距离残差继续提取下一个距离成分.

最后,将提取的距离成分进行拟线性回归,计算拟线性回归方程的回归系数,通过模型求解方法将

回归方程转换成原始变量 X 和 Y 的表达.

1.3　融入距离方差和距离相关系数的偏最小二乘(DVDCCPLS)回归算法理论流程

1.3.1　**相关定义**　假设变量 X 和 Y 的第 k 行样本分别被记为 X_k、Y_k,其中 $X \in \mathbf{R}^p$ 和 $Y \in \mathbf{R}^q$ 服从联合分布,则定义原始变量 X 和 Y 对应的距离矩阵的元素分别为 $a_{kl}=\sqrt{(X_k-X_l)^2}$,$b_{kl}=\sqrt{(Y_k-Y_l)^2}$,即距离矩阵为原始矩阵两两样本间的欧氏距离. 将距离矩阵的元素进行中心化处理得 A_{kl}、B_{kl}.

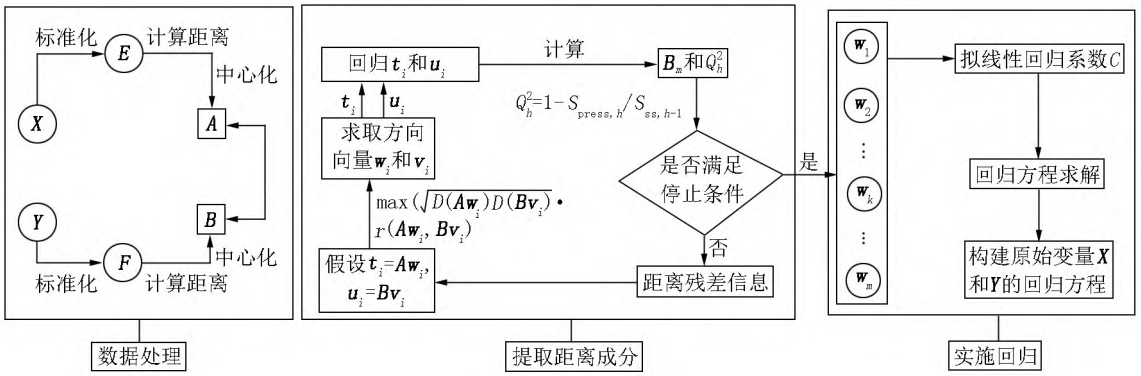

图 1　融入距离方差和距离相关系数的偏最小二乘算法结构

定义 1[18](距离方差)　假设变量 X 有 m 个样本点,则距离方差的定义为

$$V_m^2(X,X)=\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}A_{kl}^2. \tag{1}$$

定义 2[18](距离协方差)　假设变量 X 和 Y 有 m 个样本点,则距离协方差的定义为

$$V_m^2(X,Y)=\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}A_{kl}B_{kl}.$$

定义 3[18](距离相关系数)　变量 X 和 Y 之间的距离相关系数 $R_m^2(X,Y)$ 被定义为

$$R_m^2(X,Y)=\begin{cases}\dfrac{\sum_{k=1}^{m}\sum_{l=1}^{m}A_{kl}B_{kl}}{\sqrt{\sum_{k=1}^{m}\sum_{l=1}^{m}A_{kl}^2\sum_{k=1}^{m}\sum_{l=1}^{m}B_{kl}^2}}, \\ V_m^2(X,X)V_m^2(Y,Y)>0, \\ 0, V_m^2(X,X)V_m^2(Y,Y)=0.\end{cases}$$

1.3.2　**构建 DVDCCPLS 回归模型**　1) 标准化变量. 假设有自变量 $X=(x_1,x_2,\cdots,x_q)$ 和因变量 $Y=(y_1,y_2,\cdots,y_p)$,有 m 个样本点,于是构成了自变量与因变量的观测矩阵 $X'=(x_{ij})_{q\times m}$ 和 $Y'=(y_{nt})_{m\times p}$,首先将观测值进行标准化得到标准化后的矩阵 $E=(e_{ij})_{q\times m}$,$F=(f_{nt})_{m\times p}$.

2) 计算标准化矩阵两两样本间的欧氏距离并进行中心化处理. 将标准化矩阵 E、F 分别计算样本间的欧氏距离,得到距离矩阵的元素分别为 a_{kl}、b_{kl},k、$l=1,2,\cdots,m$. 将距离矩阵每个元素减去所在行的平均值,减去所在列的平均值,再加上距离矩阵 a 所有元素的平均值,最终得到中心化后的距离矩阵元素 A_{kl}、B_{kl}.

3) 令自变量为 $A=(A_1,A_2,\cdots,A_m)$,其中 A_i $(i=1,2,\cdots,m)$ 表示距离矩阵 A 的第 i 列数据. 同理,设因变量为 $B=(B_1,B_2,\cdots,B_m)$,其中 B_i $(i=1,2,\cdots,m)$ 表示距离矩阵 B 的第 i 列数据,每列有 m 个样本点.

4) 提取第 1 个距离成分. 根据偏最小二乘法的思想,提取距离成分有 2 个要求:(a) 携带原始变量信息最多;(b) 自变量与因变量之间的相关系数最大,即提取的主成分应满足

$$\begin{cases}\max(D(t_1)),\\ \max(D(u_1)),\\ \max(r(t_1,u_1)),\end{cases} \tag{2}$$

而 t_1、u_1 分别为变量 A、B 的线性组合,即

$$\begin{cases}t_1=Aw_1,\\ u_1=Bv_1,\\ \|w_1\|=1,\\ \|v_1\|=1.\end{cases} \tag{3}$$

结合式(2)和式(3)可以发现,$D(Aw_1)=w_1^{\mathrm{T}}w_1\sum_{k=1}^{m}\sum_{l=1}^{m}A_{kl}^2/m$ 和 $D(Bv_1)=v_1^{\mathrm{T}}v_1\sum_{k=1}^{m}\sum_{l=1}^{m}B_{kl}^2/m$;可以看出 $\sum_{k=1}^{m}\sum_{l=1}^{m}A_{kl}^2/m$、$\sum_{k=1}^{m}\sum_{l=1}^{m}B_{kl}^2/m$ 分别为向量 X 和 Y 的距离方差. 同理,$r(Aw_1,Bv_1)$ 采用了变量 X 和 Y 之间的距离相关系数,利用距离相关系数既可以识别变量间的线性关系,也可以识别变量间非线性关系.

利用拉格朗日算法求解参数 w_1、v_1,即

$$A^{\mathrm{T}}BB^{\mathrm{T}}Aw_1=\theta_1^2w_1,B^{\mathrm{T}}AA^{\mathrm{T}}Bv_1=\theta_1^2v_1, \tag{4}$$

其中 w_1、v_1 分别是矩阵 $A^{\mathrm{T}}BB^{\mathrm{T}}A$、$B^{\mathrm{T}}AA^{\mathrm{T}}B$ 最大特征

值对应的特征向量.

5) 分别建立 A 和 B 对 t_1,u_1 的回归方程

$$A = t_1\alpha_1^{\mathrm{T}} + A_1^*, B = u_1\beta_1^{\mathrm{T}} + B_1^*,$$

其中 $\alpha_1 = A^{\mathrm{T}}t_1/\|t_1\|^2$,$\beta_1 = B^{\mathrm{T}}u_1/\|u_1\|^2$ 是回归方程的回归系数,A_1^*、B_1^* 是方程的残差矩阵. 通过残差矩阵 A_1^*、B_1^* 取代 A、B,通过式(2) ~ 式(4) 过程求取第 2 个距离成分 t_2、u_2,不断循环,直到满足精度要求.

6) 将提取的距离成分进行回归. 提取 h 个距离成分后构建成分的回归表达式,可得

$$A = t_1\alpha_1^{\mathrm{T}} + t_2\alpha_2^{\mathrm{T}} + \cdots + t_h\alpha_h^{\mathrm{T}},$$

$$B = t_1r_1^{\mathrm{T}} + t_2r_2^{\mathrm{T}} + \cdots + t_hr_h^{\mathrm{T}} + B^*.$$

由于 $t_1,t_2,\cdots,t_h$ 可以表示成 $A_1,A_2,\cdots,A_n$ 的线性组合,所以可以得到自变量 B 和因变量 A 的回归表达式

$$B_k = u_1A_1 + u_2A_2 + \cdots + u_kA_k(k = 1,2,\cdots,m). \quad (5)$$

7) 对回归表达式进行反中心化和反标准化. 已知 A、B 是标准化后的矩阵计算欧氏距离并进行中心化处理后的矩阵,将回归方程(5) 进行反中心化和反标准化后可以得到

$$\|y_k - y_l\|_2 = \sqrt{(y_k - y_l)^2} = c_1\sqrt{(x_k - x_1)^2} + c_2\sqrt{(x_k - x_2)^2} + \cdots + c_l\sqrt{(x_k - x_l)^2}, \quad (6)$$

其中 $c_i(i = 1,2,\cdots,l)$ 为反标准化后的回归方程系数,$k,l = 1,2,\cdots,m$.

1.3.3 回归方程求解 通过 DVDCCPLS 得到的回归表达式是通过距离描述的复杂公式,它通过距离描述自变量和因变量的相关性,具体如式(6) 所示. 然而,在预测时不能直接通过式(6) 得到相应预测值,同时该公式不能直接反映原始自变量和因变量的关系. 因此,需要将距离回归方程进行求解,得到结构简洁、精度较高的回归表达式.

已知 a、b 分别是标准化后变量矩阵 E、F 的距离矩阵. 假设距离矩阵 a 和标准化变量矩阵 E 的关系、距离矩阵 b 和标准化变量矩阵 F 的关系分别为

$$E\Gamma_1 = a, b\Gamma_2 = F, \quad (7)$$

其中 Γ_1 和 Γ_2 为未知矩阵,将式(7) 的变换可以得到

$$\Gamma_1 = E^{-1}a, \Gamma_2 = b^{-1}F, \quad (8)$$

其中 E^{-1} 为标准化后矩阵 E 的逆矩阵,b^{-1} 是距离矩阵 b 的伪逆矩阵.

将式(5) 写为

$$b = \rho_l^{\mathrm{T}}a. \quad (9)$$

将式(7) 代入式(9) 得

$$F = \rho_l^{\mathrm{T}}E\Gamma_1\Gamma_2. \quad (10)$$

将式(8) 中的 $\Gamma_1 = E^{-1}a$ 代入式(10) 可得

$$F = \rho_l^{\mathrm{T}}a\Gamma_2, \quad (11)$$

其中 $\Gamma_2 = b^{-1}F$,将其转换为原始数据的表达为 $\Gamma_2 = \gamma^*(\|y_k - y_l\|_2)^{-1}y_q$,其中 $\|y_k - y_l\|_2$ 表示变量 Y 的距离矩阵,γ^* 是反标准化后的系数,Γ_2 可通过在训练数据中因变量的值计算得到.

将式(11) 展开并进行反标准化后得到最终的回归表达式为 $Y = \rho_l^{\mathrm{T}}c\|x_k - x_l\|_2\Gamma_2$.

2 实验结果及分析

2.1 数据集介绍

为了验证本文提出的 DVDCCPLS 在执行回归任务时的有效性和优势,分析 DVDCCPLS 在不同类型数据集上的适用性. 本文分别在麻杏石甘汤治疗咳嗽、哮喘和发热的中药量-效关系数据和 UCI 数据集上验证新方法的性能,并将得出的结果与其他经典的回归方法进行对比. 表 1 和表 2 分别介绍了数据集的基本信息.

表 1 麻杏石甘汤治疗咳嗽、哮喘和发热的数据集信息

数据集	特征数	因变量数	样本数
Anti-tussive	8	1	12
Anti-asthmatic	8	2	13
Anti-febrile	8	3	13

表 2 UCI 数据集信息

数据集	特征数	因变量数	样本数
Estate valuation	6	1	414
Average Localization Error	4	1	107
Gas Sensor Array1	432	2	58
Container Crane Controller	2	1	15
Gas Sensor Array2	7 501	2	928

2.2 方法评估

本文通过评价算法的回归效果和运行时间来整体评估 DVDCCPLS 的性能. 评价 DVDCCPLS 的回归效果是指评价回归后预测值和实际值的差异; 评估回归效果的指标包括均方根误差(R_{MSE}) 和绝对误差(M_{AE}).

1) 回归效果评价指标. 为了验证方法的有效性,本文采用十折交叉验证法测试模型精度,用 R_{MSE} 和 M_{AE} 作为评价回归效果的指标,计算公式分别为

$$R_{\mathrm{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, M_{\mathrm{AE}} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|,$$

其中 N 表示样本总数,y_i 为观测值,$\hat{y}_i$ 为预测值. R_{MSE}、M_{AE} 的值越小说明回归效果越好.

2) 运行时间. 运行时间是指算法开始时间与结束时间的时间差,即运行时间 = 开始时间 - 结束时间.

2.3　实验结果

将数据集在实验环境为 Window10 操作系统(64 位)、Intel(R) Core(TM) i5-7200U CPU、8 G 的 RAM 以及 Pycharm 开发平台上展开实验. 将新方法 DVDCCPLS 构建的模型采用十折交叉验证,并将实验结果与偏最小二乘法(PLS)、支持向量机(SVR)、岭回归(Ridge_reg)、决策树回归(DT_reg)和正交距离回归(ODR)进行对比,利用 R_{MSE} 和 M_{AE} 评价模型的效果. 具体实验结果如表 3 和表 4 所示,表中 y_1、y_2 和 y_3 分别表示第 1、第 2 和第 3 个因变量,最优结果以* 号标注. 另外,图 2 反映了 6 个算法经过实验后得出的 R_{MSE} 和 M_{AE} 的总体波动情况,由于实验结果的数量级差别较大,所以在绘制统计图时对 R_{MSE} 和 M_{AE} 进行归一化处理,使得所有取值在 [0,1] 之间.

结合表 3 和表 4 的实验结果和图 2 显示,DVDCCPLS 对比 PLS、SVR、Ridge_reg、DT_reg 和 ODR 在各组数据集上的实验结果可以分析如下:

1) 以 R_{MSE} 为评价指标来看,在中药量-效数据集的 6 个实验结果中,DVDCCPLS 在 5 个实验结果上表现最优; 在 UCI 数据集的实验结果中,DVDCCPLS 在 4 个结果中表现最优. 以 M_{AE} 为评价指标来看,DVDCCPLS 在 5 个中药数据的实验结果中表现最优,在 2 个 UCI 数据集的实验结果中表现最优. 综上所述,以 R_{MSE} 指标来看 DVDCCPLS 总共有 9 个结果表现最好,占总体的 69.231%; 以 M_{AE} 指标来看 DVDCCPLS 总共有 7 个结果最优,占总体的 53.846%. 这表明新方法 DVDCCPLS 在大部分数据集中表现较好,且相比其他方法有较好的回归性能. 另外,由图 2 显示,DVDCCPLS 在大部分数据集上的 R_{MSE} 和 M_{AE} 均较小,ODR 算法的 R_{MSE} 和 M_{AE} 总体趋势均比其他算法的 R_{MSE} 和 M_{AE} 更大.

2) 对于样本数较小的数据集 Anti-tussive、Anti-asthmatic、Anti-febrile、Gas Sensor Array1 和 Container Crane Controller(它们的样本数分别为 12、13、13、58 和 15),DVDCCPLS 的 R_{MSE} 和 M_{AE} 在大多数数据集的实验结果中均比其他回归算法的 R_{MSE} 和 M_{AE} 更小,仅在 Anti-asthmatic(y_1) 和 Gas Sensor Array1(y_2) 2 个数据集的实验结果中稍微比最优结果大些. 这些结果均表明新方法 DVDCCPLS 在小样本的数据集中优势较明显,这说明在计算数据的欧氏距离后再提取距离成分的方式可以更充分地利用数据中的线性和非线性信息,更好地反映原始变量的线性或非线性关系,从而提高模型的回归精度.

表 3　6 种回归算法的 R_{MSE} 实验结果对比

Datasets	DVDCCPLS	PLS	SVR	Ridge_reg	DT_reg	ODR
Anti-tussive	3.743*	4.899	3.756	9.757	5.490	31.934
Anti-asthmatic(y_1)	25.239	22.132	11.504*	58.672	22.891	131.482
Anti-asthmatic(y_2)	0.669*	0.741	0.757	1.154	0.683	1.614
Anti-febrile(y_1)	0.778*	0.924	0.868	0.951	1.100	4.590
Anti-febrile(y_2)	0.672*	0.713	1.562	0.839	1.554	24.842
Anti-febrile(y_3)	9.985*	10.135	17.689	10.454	17.761	69.857
Estate valuation	8.751*	8.780	12.413	8.897	10.210	9.167
Average Localization Error	0.228*	0.235	0.380	0.235	0.229	0.235
Gas Sensor Array1(y_1)	0.072*	0.078	0.087	0.073	0.411	117.545
Gas Sensor Array1(y_2)	0.098	0.091	0.102	0.077*	0.649	117.521
Container Crane Controller	0.127*	0.163	0.131	0.163	1.025	0.163
Gas Sensor Array2(y_1)	0.273	0.258	0.243	0.214	0.209*	4 719.195
Gas Sensor Array2(y_2)	0.311	0.267	0.233	0.208*	0.312	4 719.171

注: 带* 号表示结果最优. 下同.

表 4 6 种回归算法的 M_{AE} 实验结果对比

Datasets	DVDCCPLS	PLS	SVR	Ridge_reg	DT_reg	ODR
Anti-tussive	3.670*	4.793	3.713	9.529	5.376	4.986
Anti-asthmatic(y_1)	24.260	22.072	11.249*	57.395	22.155	10.385
Anti-asthmatic(y_2)	0.643*	0.736	0.697	1.110	0.680	1.034
Anti-febrile(y_1)	0.756*	0.898	0.834	0.875	1.087	1.933
Anti-febrile(y_2)	0.639*	0.669	1.546	0.792	1.458	4.150
Anti-febrile(y_3)	9.956*	9.374	17.062	9.840	17.105	7.702
Estate valuation	6.439	6.388	9.637	6.465	6.476	2.575*
Average Localization Error	0.173	0.173	0.297	0.173	0.173*	0.414
Gas Sensor Array1(y_1)	0.058*	0.062	0.073	0.052	1.021	9.758
Gas Sensor Array1(y_2)	0.069	0.066	0.073	0.054*	0.269	9.756
Container Crane Controller	0.122*	0.160	0.339	0.160	0.906	0.383
Gas Sensor Array2(y_1)	0.187	0.177	0.153	0.146	0.091*	56.966
Gas Sensor Array2(y_2)	0.208	0.177	0.146	0.134*	0.154	56.966

3) 维数较高的数据集 Gas Sensor Array2(其数据维数为 7 501),该数据集有 2 个因变量,对应的数据分别为 Gas Sensor Array2(y_1) 和 Gas Sensor Array2(y_2). 实验结果表明,与这 2 个数据集的最优实验结果相比,DVDCCPLS 的 R_{MSE} 仅分别相差 0.064 和 0.103,M_{AE} 仅分别相差 0.096 和 0.074. 这也说明 DVDCCPLS 对于高维数据也同样适用. 回归精度略微差的原因是: DVDCCPLS 在数据处理过程中是计算两两样本间的欧氏距离,且样本间的欧氏距离是在所有变量的任意 2 个取值相减取平方后累加再开根号的过程. 当变量的维数过高时,得出的距离矩阵将会产生较大的误差,这将导致信息的损失,并且在回归方程的求解过程中会产生模型的 2 次误差. 因此,当数据维数较高时,DVDCCPLS 的回归误差高于其他回归方法的回归误差.

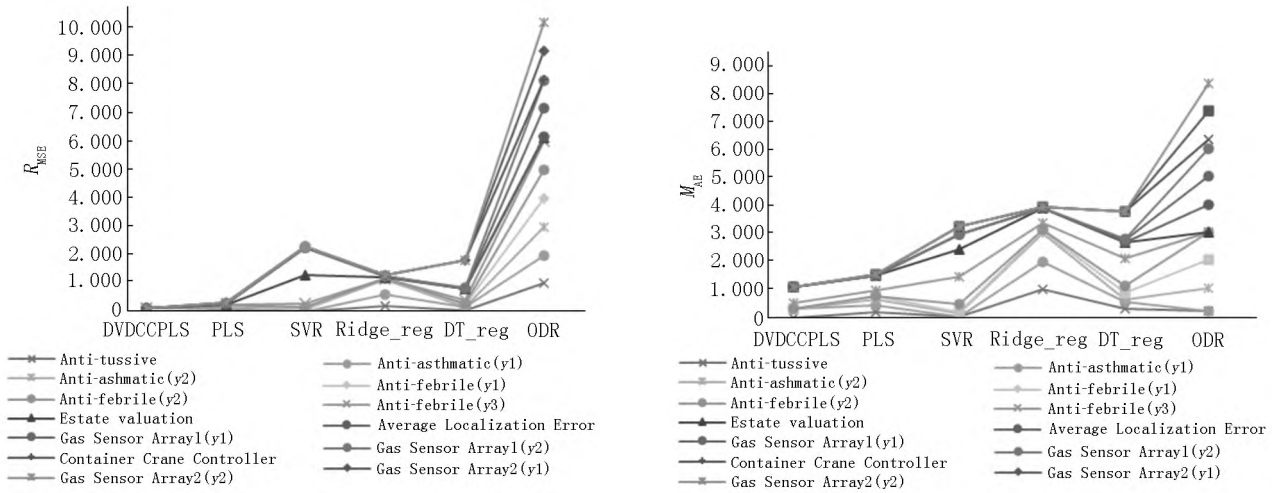

图 2 6 种算法对不同数据集的实验结果对比

另外,为了研究算法 DVDCCPLS 在执行回归任务时的时间复杂度,本文对 DVDCCPLS 在各 UCI 数据集上的运行时间进行了分析,并与传统 PLS 算法的时间复杂度进行对比. 为了公平比较,2种方法均运行在相同的软硬件环境中,同时 DVDCCPLS 和 PLS 方法均采用十折交叉对模型进行验证,时间越少说明计算时间复杂度越低. 具体实验结果如表 5 所示.

表 5　算法的运行时间结果对比

数据集	样本维数	样本数	运行时间/s	
			PLS	DVDCCPLS
Estate valuation	7	414	28.971*	71.566
Average Localization Error	5	107	4.640	3.843*
Gas Sensor Array1	434	58	10.158	3.168*
Container Crane Controller	3	15	0.208	0.180*
Gas Sensor Array2	7 503	928	4 974.490	314.292*

根据表 5 的运行时间对比分析可知,对于 Average Localization Error、Gas Sensor Array1、Container Crane Controller 和 Gas Sensor Array2 这 4 组数据集,DVDCCPLS 的运行时间比 PLS 的运行时间更短,它们分别相差 0.797、6.990、0.028 和 4 660.198 s. 值得注意的是: 对于数据集 Gas Sensor Array2,2 种方法相差悬殊,DVDCCPLS 的运行时间比 PLS 的运行时间少 4 660.198 s; 对于数据集 Estate valuation,PLS 的运行时间比 DVDCCPLS 的运行时间少 42.595 s.

由运行时间的分析可以看出,DVDCCPLS 在维数较高的数据集上的运行时间大大降低. 其原因是: DVDCCPLS 在提取距离成分过程时的时间复杂度与样本数量有关,而与样本维数无关. 该方法为高维小样本数据集的回归提供可行性,可以用较小的时间复杂度对高维数据实施回归; 当样本数较大时,DVDCCPLS 的运行时间将大大提高.

3　结论

由于 PLS 在基于 Pearson 系数提取主成分以及使用多元线性回归进行主成分回归中存在一些不足,本文提出一种新的回归方法 DVDCCPLS. 新方法通过距离方差反映原数据信息,采用距离相关系数度量自变量与因变量之间的相关性,以及采用拟线性回归方法对距离成分进行回归. 本文从多个角度进行实验设计,分别在中药量-效关系数据和 UCI 数据集上进行验证. 首先,通过不同指标和不同方法的对比来评价新方法的回归效果; 其次,分析 DVDCCPLS 方法在不同数据集上的优势和不足及其原因; 最后对比了 DVDCCPLS 和 PLS 的时间复杂度. 结果表明新方法对于线性和非线性数据均具有更好的预测性能. 同时根据实验效果值得注意的是,DVDCCPLS 对于小样本数据效果更好,因此该方法可以较好地运用于小样本的数据集中. 接下来,将继续优化算法,进一步研究该算法在高维、大样本数据中的应用,以及提高模型的精确度.

4　参考文献

[1] GELADI P,KOWALSKI B R. Partial least squares regression: a tutorial [J]. Analytica Chimica Acta,1986,185: 1-17.

[2] YOU Xinge,MOU Yi,YU Shujian,et al. Mixed-norm partial least squares [J]. Chemometrics and Intelligent Laboratory Systems,2016,152: 42-53.

[3] MALTHOUSE E C,TAMHANE A C,MAH R S H. Nonlinear partial least squares [J]. Computers & Chemical Engineering,1997,21(8): 875-890.

[4] LIU Hongbin,YANG Chong,BENGT C,et al. Dynamic nonlinear partial least squares modeling using Gaussian process regression [J]. Industrial & Engineering Chemistry Research,2019,58(36): 16676-16686.

[5] 尚栋,孙兰香,齐立峰,等. 基于循环变量筛选非线性偏最小二乘的 LIBS 铁矿浆定量分析 [J]. 中国激光,2021,48(21): 171-179.

[6] MA Hao,WANG Yan,JI Zhicheng. A novel dynamic nonlinear partial least squares based on the cascade structure [J]. International Journal of Robust and Nonlinear Control,2022,32(6): 3584-3605.

[7] 贾润达,毛志忠,王福利. 基于 KPLS 模型的间歇过程产品质量控制 [J]. 化工学报,2013,64(4): 1332-1339.

[8] JIAO Jianfang,ZHAO Ning,WANG Guang,et al. A nonlinear quality-related fault detection approach based on modified kernel partial least squares [J]. ISA Transactions,2016,66: 275-283.

[9] ZHU Bao,CHEN Zhongsheng,HE Yanlin,et al. A novel nonlinear functional expansion based PLS (FEPLS) and its soft sensor application [J]. Chemometrics and Intelligent Laboratory Systems,2017,161: 108-117.

[10] WANG Yanxia,CAO Hui,ZHOU Yan,et al. Nonlinear partial least squares regressions for spectral quantitative analysis [J]. Chemometrics and Intelligent Laboratory Systems,2015,148: 32-50.

[11] 鲁庆华,任康乐,周凤玺. 基于偏最小二乘法实现非线性回归分析 [J]. 甘肃科技,2005,21(11): 146-148.

[12] MERINO A,GARCIA-ALVAREZ D,SAINZ-PALMERO G,et al. Knowledge based recursive non-linear partial least squares (RNPLS) [J]. ISA Transactions,2020,100:481-494.

[13] 李雄威,郭晓雅,李庚达,等. 一种基于非线性偏最小二乘的风电机组齿轮箱状态监测方法 [J]. 可再生能源,2022,40(10):1346-1351.

[14] LAVOIE F B,MUTEKI K,GOSSELIN R. A novel robust NL-PLS regression methodology [J]. Chemometrics and Intelligent Laboratory Systems,2018,184:71-81.

[15] 谢文龙. 三次样条函数的构造方法 [J]. 江南学院学报,2000,15(2):90-93.

[16] INDAHL U. A twist to partial least squares regression [J]. Journal of Chemometrics,2005,19(1):32-44.

[17] LAVOIE F B,MUTEKI K,GOSSELIN R. Generalization of powered-partial-least-squares [J]. Chemometrics and Intelligent Laboratory Systems,2018,179:1-11.

[18] PENG Shan,PENG Silong,TANG Liang,et al. A nonlinear partial least squares with slice transform based piecewise linear inner relation [J]. Chemometrics and Intelligent Laboratory Systems,2015,143:97-110.

[19] LI Fachao,YANG Kuo. Research of the regression method based on quasi-linear function [EB/OL]. [2022-08-13]. https://ieeexplore.ieee.org/document/5662989/.

[20] WOLD S,RUHE A,WOLD H,et al. The collinearity problem in linear regression: the partial least squares (PLS) approach to generalized inverses [J]. SIAM Journal on Scientific and Statistical Computing,1984,5(3):735-743.

[21] SZÉKELY G J,RIZZO M L,BAKIROV N K. Measuring and testing dependence by correlation of distances [J]. Annals of Statistics,2007,35(6):2769-2794.

The Regression Method of PLS Fusing Distance Variance and Distance Correlation Coefficient

NIE Bin,DU Yuwen,DU Jianqiang*,ZHANG Yuchao,ZHENG Xuepeng,JIN Haike
(School of Computer,Jiangxi University of Chinese Medicine,Nanchang Jiangxi 330004,China)

Abstract: Partial least square (PLS) internally adopts Pearson coefficient to measure the correlation between independent and dependent variables,but the extracted components cannot ensure the strongest interpretation. In addition,PLS applies linear regression to the extracted components,which is unable to reflect the functional relationship between variables truly. Therefore,the regression method of partial least square fusing distance variance and distance correlation coefficient (DVDCCPLS) is proposed to solve the above problems. DVDCCPLS extracts the distance component based on the distance variance and distance correlation coefficient,and then performs quasilinear regression to obtain the distance regression equation. Finally,the distance regression equation is converted into the expression of the original data by the model solution method,and the regression model with simple structure and high precision is obtained in the end. The Maxingshigan decoction datasets and UCI datasets are respectively used to test the performance of DVDCCPLS,and the other five classical regression algorithms are compared with DVDCCPLS. The results show that DVDCCPLS has better regression effects and performances.

Key words: partial least square; distance variance; distance correlation coefficient; regression equations; quasilinear

(责任编辑: 冉小晓)