低秩特征选择多输出回归算法

杨利锋[1,2],林大华[3],邓振云[1,2],李永钢[1,2]
YANG Lifeng[1,2], LIN Dahua[3], DENG Zhenyun[1,2], LI Yonggang[1,2]

1.广西多源信息挖掘与安全重点实验室,广西 桂林 541004
2.广西区域多源信息集成与智能处理协同创新中心,广西 桂林 541004
3.广西电化教育馆,南宁 530021
1.Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi 541004, China
2.Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, Guilin, Guangxi 541004, China
3.Guangxi Center for Educational Technology, Nanning 530021, China

YANG Lifeng, LIN Dahua, DENG Zhenyun, et al. Low-rank feature selection for multiple-output regression algorithm. Computer Engineering and Applications, 2017, 53(20):116-121.

Abstract: To solve the issue of the existing regression models do not well take advantage of the correlation between inputs and outputs, and among outputs, also between samples, to take the multiple output regression analysis for high-dimensional data, it proposes a novel multiple output regression method called Low-rank Feature Selection for Multiple-output Regression algorithm(for short LFS_MR). The method can catch the correlation structures of outputs via a low-rank regression model with a low-rank constraint. Specially, it is innovative that the method conducts sample selection via an $L_{2,p}$-norm on this low-rank regression model, which can avoid the interference of noise and outliers reasonably. What's more, the method conducts feature selection by applying an $L_{2,p}$-norm regularization term to penalty the regression coefficient matrix, which handles with the correlations between inputs and outputs efficiently, and solves the problem of curse of dimensionality for the high-dimensional data. The experimental results on many realistic datasets show that the proposed method can obtain very good results when conduct a multiple output regression analysis for high-dimensional data.
Key words: multiple-output regression; low-rank regression; regression coefficient matrix; feature selection

摘 要:针对现有回归算法没有考虑利用特征与输出的关系,各输出之间的关系,以及样本之间的关系来处理高维数据的多输出回归问题易输出不稳定的模型,提出一种新的低秩特征选择多输出回归方法。该方法采用低秩约束去构建低秩回归模型来获取多输出变量之间的关联结构;同时创新地在该低秩回归模型上使用$L_{2,p}$-范数来进行样本选择,合理地去除噪音和离群点的干扰;并且使用$L_{2,p}$-范数正则化项惩罚回归系数矩阵进行特征选择,有效地处理特征与输出的关系和避免"维灾难"的影响。通过实际数据集的实验结果表明,提出的方法在处理高维数据的多输出回归分析中能获得非常好的效果。
关键词:多输出回归;低秩回归;回归系数矩阵;特征选择
文献标志码:A **中图分类号:**TP181 doi:10.3778/j.issn.1002-8331.1604-0366

基金项目:国家自然科学基金(No.61450001,No.61263035,No.61573270);国家高技术研究发展计划(863)(No.2012AA011005);国家重点基础研究发展规划(973)(No.2013CB329404);中国博士后科学基金(No.2015M570837);广西自然科学基金(No.2012GXNSFGA060004,No.2014jjAA70175,No.2015GXNSFAA139306,No.2015GXNSFCB139011);广西八桂创新团队、广西百人计划和广西高校科学技术研究重点项目(No.2013ZD04);广西研究生教育创新计划项目(No.YCSZ2016046,No.YCSZ2016045)。
作者简介:杨利锋(1989—),男,硕士,主要研究领域为数据挖掘、机器学习;林大华(1979—),男,通讯作者,高级教师,主要研究领域为数据挖掘,E-mail:517567113@qq.com;邓振云(1991—),男,硕士,主要研究领域为数据挖掘、机器学习;李永钢(1989—),男,硕士,主要研究领域为数据挖掘、机器学习。
收稿日期:2016-04-26 **修回日期:**2016-07-06 **文章编号:**1002-8331(2017)20-0116-06
CNKI**网络优先出版:**2016-09-29, http://www.cnki.net/kcms/detail/11.2127.TP.20160929.1618.006.html

1 引言

一直以来,多输出回归分析是机器学习和统计学的重点研究问题。这种分析方法根据数据的多个特征值预测多个输出实值,目前,在很多领域里有应用,如预测植被多个生长状况的生态系统,从遥感图像中评估多个生物物理参数,用多个宏观经济变量和以往的股价来预测股票的价格等。

传统的线性回归[1]模型能解多输出回归问题,其将多个特征值作为输入分别学习与各个输出变量相应的回归系数,获得回归模型。然而,当输出变量很多时,输出之间往往会存在某种关联结构[2],该模型没有利用这种关联结构。后来,为了利用该关联结构改善所学习的回归模型,Anderson[3]首先提出了一种低秩回归方法,用一个低秩来约束回归系数矩阵而获得多输出变量之间的关联结构。此后,低秩回归模型受到研究者热点关注和研究。研究者们[4]通过迹范数正则化项来学习低秩回归系数矩阵,获得多输出变量之间的低秩结构。但是,该方法不能明确地选择和调节所学到的回归系数矩阵的秩的大小。通过两个低秩矩阵的乘积表示回归系数矩阵,巧妙地将回归系数的秩约束为小于特征维数和输出变量数这个条件融入到低秩回归模型中,能明确地决定该秩的大小。但是,该方法没有进行属性约减,对高维数据进行回归分析的效果有限。实际数据往往拥有大量的特征维数[5],如基因数据,图像数据,文本数据等。为了提高对这些高维数据回归分析的效果,通常需要收集大量的样本来学习模型。然而,由于各种原因很难获得足够的样本,且还会遇到“维灾难”问题[6]。为此,研究者们[7]在低秩回归模型的基础上,增加特征选择的处理步骤,即同时使用属性约减的方法处理高维数据。但是,这种方法没有对样本进行选择,可能会受噪音和离群点的干扰。

因此,为了更好地处理高维数据多输出回归问题,本文方法首先通过低秩约束来构建低秩回归模型,同时在该模型上使用 $\ell_{2,p}-norm$ 进行样本选择,且利用 $\ell_{2,p}-norm$ 正则化项惩罚回归系数矩阵,获得全局最优的回归系数矩阵。接着,依据该回归系数矩阵对训练样本进行特征选择得到新训练样本集。然后,在libSVM内进行5-折交叉验证来学习回归模型。最后,在10-折交叉验证的框架上进行以上步骤,获得最好的回归模型。在构建回归系数矩阵的过程中,低秩回归可确保输出变量之间的关联结构得到使用,充分利用了当实际数据具有大量的输出变量时其相互之间会存在某种关联结构的本质;低秩回归上的 $\ell_{2,p}-norm$ 能够对样本进行选择,有效地去除噪音和离群点的干扰,提高回归分析的预测效果; $\ell_{2,p}-norm$ 正则化项惩罚回归系数矩阵而对特征进行选择,能利用特征与输出之间的关系,有利于对高维数据进行回归分析。本文将这种多输出回归方法称为低秩特征选择多输出回归算法简称为LFS_MR(Low-rank Feature Selection for Multiple-output Regression algorithm)。

2 相关理论

给定一个训练数据集合 $D=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\}=(\boldsymbol{X},\boldsymbol{Y})\in\boldsymbol{R}^{n\times(d+c)}$,其中 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ 表示一个样本点, $\boldsymbol{x}_i\in\boldsymbol{R}^{1\times d}(i=1,2,\cdots,n)$ 表示具有 d 维特征的输入向量, $\boldsymbol{y}_i\in\boldsymbol{R}^{1\times c}(i=1,2,\cdots,n)$ 表示具有 c 个输出的输出向量, n 为样本数, $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n]^{\mathrm{T}}\in\boldsymbol{R}^{n\times d}$ 为输入矩阵, $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_i,\cdots,\boldsymbol{y}_n]^{\mathrm{T}}\in\boldsymbol{R}^{n\times c}$ 为输出矩阵(或响应矩阵)。多输出回归[8](multiple-output regression)的目的是指数据集 D 中寻找一个函数模型 h,使得 $\boldsymbol{X}$ 映射到相应的 $\boldsymbol{Y}$,即

$$\boldsymbol{x}_i\xrightarrow{h}\boldsymbol{y}_i$$

应用到高维数据的回归分析中,即希望充分考虑样本特征与各输出之间的关系,从中找到一个投影矩阵 $\boldsymbol{W}$,使得 $\boldsymbol{Y}$ 与 $\boldsymbol{XW}$ 之间的差值尽可能小,这一般可通过最小二乘损失函数实现,即

$$\min_{\boldsymbol{W}}\|\boldsymbol{Y}-\boldsymbol{XW}\|_F^2 \tag{1}$$

其中, $\boldsymbol{W}\in\boldsymbol{R}^{d\times c}$ 为回归系数矩阵,其解为: $\boldsymbol{W}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$。值得注意的是,该方法只是简单地将多输出问题转换为多个独立的单输出问题,然后对每个单输出进行最小二乘估计,而没有考虑到各输出之间的潜在关联。实际应用中已经有证明:多输出回归相比于单输出回归可以产生一个更好的预测性能。因此为了实现同时预测多个输出变量,充分利用了各输出变量之间的关联[9],对回归系数矩阵 $\boldsymbol{W}$ 进行如下低秩约束:

$$\min_{\boldsymbol{W}}\|\boldsymbol{Y}-\boldsymbol{XW}\|_F^2,\ \text{s.t. } rank(\boldsymbol{W})\leqslant\min(d,c) \tag{2}$$

从公式(2)中可以看出,该低秩约束明显地减少了需要计算的参数,提高了最小二乘估计的效率。此外,在实际数据中,噪音和离群点往往会增大系数矩阵的秩[10]。因此,在将高维数据投影到低维结构的过程中,对式(1)中进行低秩约束也是合理的。在对公式(1)进行低秩约束的过程中,回归系数矩阵 $\boldsymbol{W}$ 可以表示成两个秩不大于 r 的矩阵的乘积,即

$$\boldsymbol{W}=\boldsymbol{BA}^{\mathrm{T}} \tag{3}$$

其中, $\boldsymbol{B}\in\boldsymbol{R}^{d\times r}$ 为回归系数矩阵, $\boldsymbol{A}\in\boldsymbol{R}^{c\times r}$ 为子空间学习系数矩阵。通过式(3)进行替代,可以得到低秩线性回归模型的一般形式:

$$\min_{\boldsymbol{A},\boldsymbol{B}}\left\|\boldsymbol{Y}-\boldsymbol{XBA}^{\mathrm{T}}\right\|_F^2 \tag{4}$$

3 LFS_MR算法

根据公式(2)可知,采用低秩约束去构建的低秩回归模型可以很好地利用各输出变量之间的关联结构。

但是在处理海量高维数据时,仍然会有大量的特征和样本对回归的预测是无用的。因此,为了提高回归分析的准确率,对目标函数(4)实行进一步优化,即对已获得的低秩回归模型进行特征选择和样本选择,排除低秩约束过程中存在的冗余属性以及噪音和离群点样本,提高回归分析的效率。

为了能够去除数据中的噪音和离群点样本,通常会对数据集进行样本选择,即对于数据集里的每一个训练样本,选择与其线性相关的样本进行表示,即,利用 $\ell_{2,1}-norm$ 去替代公式(4)中的 ℓ_F-norm 去惩罚 $(Y-XBA^{\mathrm{T}})$ 的所有行系数。但是,实际应用[11]表明, $\ell_{2,p}-norm$ 可以通过调节参数 p 而能更好地控制样本之间相关性结构的程度,因此,通过最小化 $\ell_{2,p}-norm$ 去压缩回归矩阵 BA^{T} ,排除离群点和不相干样本的干扰,即

$$\min_{A,B}\left\|Y-XBA^{\mathrm{T}}\right\|_{2,p} \tag{5}$$

其中 $0<p<2$ 。当 $p=1$ 时,即为标准的 $\ell_{2,1}-norm$;当对 p 进行调参时,可以控制矩阵 BA^{T} 中非零行的个数。因此,对于排除冗余和不相关样本而言, $\ell_{2,p}-norm$ 具有更高的效率。但是,目标函数(5)是凸函数,易知其解为 $BA^{\mathrm{T}}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ 。而在实际应用中 $X^{\mathrm{T}}X$ 不一定可逆。为此,考虑引入一个 $\ell_{2,p}-norm$ 正则化项使其可逆,并且对数据集中的冗余和不相关特征进行去除,另外还考虑对子空间学习系数矩阵 A 引入一个正交约束,去充分考虑各输出变量之间的相关性。最后,得到如下目标函数:

$$\min_{A,B}\left\|Y-XBA^{\mathrm{T}}\right\|_{2,p}+\alpha\|B\|_{2,P},\ \text{s.t.}\ A^{\mathrm{T}}A=I_r \tag{6}$$

其中 $\|B\|_{2,P}=\left[\sum_{i=1}^{d}\left(\sum_{j=1}^{r}\left|B_{ij}\right|^2\right)^{p/2}\right]^{1/p}$,α 为非负调整参数。正交旋转约束条件 $A^{\mathrm{T}}A=I_r\in R^{r\times r}$,为正交约束。 $\ell_{2,p}-norm$ 通过惩罚回归系数矩阵 B 的行系数,对整个低秩回归模型进行属性选择。使得对每个测试样本都通过相关性高的特征进行预测。而对于子空间系数矩阵 A 采用正交约束,则可以很好地确保每个输出都是由共同的特征进行预测,同时也确保了低秩回归约束可以通过考虑各输出之间的关系来进行子空间选择。

根据以上分析可知,本文提出的目标函数(6),不仅通过低秩约束 $rank(\boldsymbol{W})=r\leqslant\min(d,c)$ 构造低秩回归模型,合理地利用了输出之间的关联,还同时在该低秩回归模型上使用 $\ell_{2,p}-norm$ 从各个输出之间的关联方面对样本进行选择,去除了数据中离群点的干扰。而且,使用 $\ell_{2,p}-norm$ 正则化项惩罚回归系数矩阵 B ,从特征输入与输出之间的关系方面出发进行特征选择,寻找高维数据中的低维结构,有利于进行高维数据的回归分析。因此,本文提出的方法能够结合利用特征输入与输出之间关系,各输出之间的关系,以及样本之间的关系来构建多输出回归模型。该模型通过低秩回归和特征选择过程而能够非常好地处理高维数据的多输出回归分析,并且可以去除离群点对构建一个良好模型的干扰。

最后,给出本文提出的LFS_MR算法的具体实现步骤如下:

算法 1 LFS_MR算法

输入 数据集 $D=(X,Y)\in R^{n\times(d+c)}$,参数 α,p,r

输出 aCC和aRMSE的均值

(1)对数据集 D 进行规范化处理和10-折交叉验证。

(2)通过算法2获得最优化的 A,B 。

(3)通过 B 对训练样本进行特征选择,得到新训练集。

(4)在libSVM中进行5-折交叉验证学习模型,求出测试样本的预测值。

(5)计算aCC和aRMSE并获得均值。

传统线性回归模型[1]通过求解最小二乘函数 $\min_{\boldsymbol{W}}\|Y-X\boldsymbol{W}\|_F^2$,以获得使预测输出 $X\boldsymbol{W}$ 与真实输出 Y 的差值最小的回归系数矩阵 $\boldsymbol{W}$,其中 X 是输入数据。求解该函数得: $\boldsymbol{W}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$,从此式可知,该方法仅利用输入与输出之间的关系,并没有考虑多输出之间的关联。为此,本文方法通过秩被约束为 r 的两个矩阵的乘积表示 $\boldsymbol{W}$,即 $\boldsymbol{W}=BA^{\mathrm{T}},rank(A)=r,rank(B)=r,$ 构造低秩回归模型从而能利用多输出之间的关联。

而通过迹范数寻找多输出之间关联的方法[4],其目标函数为 $\min_{\boldsymbol{W}}\|Y-X\boldsymbol{W}\|_F^2+\lambda\|\boldsymbol{W}\|_*$ 。因此,可知该方法通过调节参数 λ ,只是能调整 $\boldsymbol{W}$ 的秩的大小,然而,并不能明确地决定 $\boldsymbol{W}$ 的秩的大小。有研究表明[7],在实际应用中明确地选择 $\boldsymbol{W}$ 的秩的大小能获得更好的模型预测效果。为此,本文方法通过确定的 r ,约束 $rank(A)=rank(B)=r\leqslant\min(d,c)$,其中 d 表示输入数据的维数, c 表示输出的维数,从而将 $\boldsymbol{W}$ 的秩明确地约束为 r ,即 $rank(\boldsymbol{W})=rank(BA^{\mathrm{T}})=r$ 。

4 目标函数优化

本文采用迭代算法来优化目标函数(6),主要的迭代过程有以下两步:(1)固定 A 更新 B 和(2)固定 B 更新 A 这两个步骤,直到目标函数收敛到最优解,其具体实现介绍如下:

(1)固定 A 更新 B

由于子空间矩阵 A 存在正交约束 $A^{\mathrm{T}}A=I$,因此可对目标函数(6)的低秩回归模型进行如下操作:

$$\left\|Y-XBA^{\mathrm{T}}\right\|_{2,p}=\left\|(Y-XBA^{\mathrm{T}})(A,A'')\right\|_{2,p}=\left\|YA-XB\right\|_{2,p}+\left\|YA''\right\|_{2,p} \tag{7}$$

式(7)中 $\|YA''\|_{2,p}$ 不含矩阵 B ,因此,当 A 固定时,对目

标函数(6)的优化问题可转换为对如下目标函数的优化：

$$\min_B \|YA-XB\|_{2,p}+\alpha\|B\|_{2,P} \tag{8}$$

对目标函数(8)进一步推导可得：

$$\min_B \mathrm{tr}[(YA-XB)^{\mathrm{T}}N(YA-XB)]+\alpha\mathrm{tr}(B^{\mathrm{T}}QB) \tag{9}$$

其中，α 为调优参数，$N\in R^{n\times n}$ 和 $Q\in R^{r\times r}$ 都是对角矩阵，并且它们的主对角元素分别为 $N_{ii}=\dfrac{1}{\frac{2}{p}\|(YA-XB)_i\|_2^{2-p}}$，$i=(1,2,\cdots,n)$ 和 $Q_{jj}=\dfrac{1}{\frac{2}{p}\|B_j\|_2^{2-p}}(j=1,2,\cdots,r)$，然后对式(9)中 B 的每一行求导，并令其等于0，则可得：

$$B=(X^{\mathrm{T}}NX+\alpha Q)^{-1}X^{\mathrm{T}}NYA \tag{10}$$

(2)固定 B 更新 A

通过固定 B，目标函数(6)可简化为如下形式：

$$\min_A \|Y-\tilde{X}A^{\mathrm{T}}\|_{2,p},\ \text{s.t. } A^{\mathrm{T}}A=I_r \tag{11}$$

其中，$\tilde{X}=XB\in R^{n\times r}$。可知目标函数(11)事实上是一个正交普鲁克(Procrustes)问题，通过对 $Y^{\mathrm{T}}\tilde{X}=USV^{\mathrm{T}}$ 进行奇异值分解，可知子空间矩阵 A 的最优解为 UV^{T}，其中 $U\in R^{c\times r}$，$V\in R^{r\times r}$。

综上通过固定 A 更新 B，和固定 B 更新 A，对 A 和 B 进行迭代优化，使得目标函数(6)在每次迭代中都单调递减，直至收敛到最优解。最后，给出迭代[12-14]优化算法2的伪代码，如下：

算法2 目标函数优化方法

输入 $X\in R^{n\times d}, Y\in R^{n\times c}, \alpha, p, r$

输出 A, B

(1)初始化迭代次数 $t=0$。

(2)初始化 $A(0)$ 为一个随机对角矩阵，其中 $A(t)$ 表示 A 的第 t 次迭代的结果。

(3)while(相邻两次迭代的目标函数值的差大于 10^{-5} 且 $t<$ 总迭代次数)

{

① 通过式(10)，计算 $B(t+1)$；

② 通过式(11)，计算 $A(t+1)$；

③ 通过公式：

$$N_{ii}=\frac{1}{\frac{2}{p}\|(YA-XB)_i\|_2^{2-p}}, i=(1,2,\cdots,n)$$

计算 $N(t+1)$；

④ 通过公式：

$$Q_{jj}=\frac{1}{\frac{2}{p}\|B_j\|_2^{2-p}}, j=(1,2,\cdots,r),$$ 计算 $Q(t+1)$；

⑤ $t=t+1$；

}

结束。

5　实验分析与讨论

5.1　实验介绍

本文算法通过MATLAB语言编程，且所有实验都是在Win7系统下的MATLAB 2014软件上运行测试。

本文实验中用到的数据集为EDM[15]，和来自于多输出数据集[16]的ATP1d，OES97，ATP7d(已对数据集的outputs进行了标准化处理)，数据集的基本情况如表1所示。

表1　实验数据集的基本信息

dataset	samples	Features	Outputs
EDM	154	16	2
ATP1d	337	411	6
ATP7d	296	411	6
OES97	334	263	16

与本文方法对比的算法有低秩线性回归[7]算法(Low-Rank Linear Regression，LRLR)、低秩岭回归算法[7](Low-Rank Ridge Regression，LRRR)、稀疏低秩回归算法[7](Sparse Low-Rank Regression，SLRR)、稀疏多任务回归和特征选择算法[17](Sparse MultitAsk Regression and feature selection，SMART)，以及新图结构稀疏算法[18](LSG21)。

为了比较LFS_MR算法与对比算法的性能，本文用均方差误差(average Root Mean Square Error，aRMSE)以及平均相关系数(average Correlation Coefficient，aCC)作为评价指标。其中，多次实验结果的均方差误差可以反映算法的稳定性，其值越小，稳定性越好。而相关系数能反映预测值与测试样本目标值之间的相关性，其值越大，预测值越接近目标值，即预测的可信度越强。均方差误差和平均相关系数的计算公式[8]分别如下：

$$aRMSE=\frac{1}{d}\sum_{i=1}^{d}\sqrt{\frac{\sum_{j=1}^{ntest}(y_i^{(j)}-\hat{y}_i^{(j)})^2}{ntest}}$$

$$aCC=\frac{1}{d}\sum_{i=1}^{d}\frac{\sum_{j=1}^{ntest}(y_i^{(j)}-\bar{y}_i)(\hat{y}_i^{(j)}-\bar{\hat{y}}_i)}{\sqrt{\sum_{j=1}^{ntest}(y_i^{(j)}-\bar{y}_i)^2\sum_{j=1}^{ntest}(\hat{y}_i^{(j)}-\bar{\hat{y}}_i)^2}}$$

其中 $ntest$ 表示测试样本数，y_i 为目标值，$\hat{y}_i$ 为预测值，$\bar{y}_i$ 为目标值均值，$\bar{\hat{y}}_i$ 为预测值均值。

实验采用10-折交叉验证以比较所有的方法。其首先将整个数据集随机分成均匀的10份，然后选择其中的1份作为测试集，其余的9份作为训练集。并且在该10-折交叉验证里面进行5-折交叉验证用libSVM训练来学习回归模型。对于参数模型的选择，本文设置调制参数 $\alpha\in\{10^{-7},10^{-6},\cdots,10^{7}\}$，回归系数矩阵的秩 $r\in\{1,2,\cdots,\min(d,c)\}$ 和 $\ell_{2,p}-norm$ 中的 $p\in\{0.1,0.2,\cdots,1.9\}$，以及libSVM工具箱中的 $(c,g)\in\{2^{-5},2^{-4},\cdots,2^{5}\}$。

5.2 实验结果与分析

本文实验在10-折交叉验证的框架内进行libSVM内部的5-折交叉验证,每次返回5-折交叉验证的最好aCC结果和aRMSE结果,所以10-折交叉验证可分别获得10个aCC结果和aRMSE结果,然后取其10个结果的均值用作比较。其中,LRLR,LRRR,SLRR,SMART,LSG21和本文提出的LFS_MR方法在上述数据集上实验的aCC结果如表2。

表2 所有对比算法在四个数据集上的aCC结果

算法	EDM	OES97	ATP7d	ATP1d
LRLR	0.809 7	0.901 8	0.889 3	0.922 0
LRRR	0.805 1	0.902 0	0.889 9	0.922 9
SLRR	0.805 1	0.902 1	0.890 2	0.923 3
SMART	0.810 5	0.901 4	0.887 1	0.920 1
LSG21	0.804 8	0.902 8	0.890 7	0.923 3
LFS_MR	**0.818 3**	**0.904 2**	**0.941 7**	**0.944 2**

从表2数据可知,LFS_MR算法在EDM,OES97,ATP7d和ATP1d数据集上的平均相关系数结果均高于LRLR,LRRR,SLRR,SMART和LSG21算法的平均相关系数。其中,LFS_MR算法在ATP1d数据集上取得了最优的效果,其平均相关系数为94.42%,相比LRLR,LRRR,SLRR,SMART和LSG21算法分别高于2.22%,2.13%,2.09%,2.41%和2.09%。另外,在ATP7d数据集上,LSF_MR算法的aCC结果为94.17%,远高于对比算法LRLR,LRRR,SLRR,SMART和LSG21的aCC,分别高于5.24%,5.18%,5.15%,5.46%,5.10%。本文的LFS_MR算法在EDM数据集和OES97数据集上同样能够取得很好的aCC结果。

在这四个数据集上实验的所有对比算法也都获得很好的aCC结果。其中,LRLR算法通过低秩约束可以利用多输出变量之间的关联,而没有进行特征选择,所以在高维数据的多输出回归中获得的aCC结果略低于其他算法的结果;LRRR和SLRR算法均通过低秩约束去构建低秩回归模型,并且进行特征选择,其中LRRR利用 ℓ_F-norm 惩罚回归系数矩阵来进行特征选择,而SLRR则利用 $\ell_{2,1}-norm$ 正则化项惩罚回归系数矩阵来进行特征选择,但是这两种方法都没有对样本进行选择,数据中的噪音和离群点会干扰回归模型的学习,所以其获得的aCC结果均比LSF_MR的aCC结果差。

而SMART算法利用 $\ell_{2,1}-norm$ 项对高维数据的特征进行选择,并且使用 ℓ_1-norm 确保系数矩阵得到稀疏。但是,该方法没有利用输出变量之间的关联结构,所以获得的aCC结果比LRRR和SLRR以及LSG21的aCC结果低;LSG21算法利用图结构稀疏的方法获得多输出回归的结果,但是没有对样本进行选择,并且在构建图结构的过程中计算代价很大,不利于大量数据的多输出回归分析。

另外,LRLR,LRRR,SLRR,SMART,LSG21和本文提出的LFS_MR方法在上述四个数据集上实验的aRMSE结果如表3。

表3 所有对比算法在四个数据集上的aRMSE结果

算法	EDM	OES97	ATP7d	ATP1d
LRLR	0.045 3	0.016 3	0.006 7	0.007 0
LRRR	0.047 4	0.016 3	0.006 7	0.007 0
SLRR	0.047 4	0.016 2	0.006 6	0.006 9
SMART	0.047 5	0.012 9	0.006 7	0.007 1
LSG21	0.047 4	0.012 8	0.006 7	0.006 9
LFS_MR	**0.045 0**	**0.012 8**	**0.006 5**	**0.006 8**

从表3的aRMSE实验结果可知,LFS_MR算法在数据集EDM,OES97,ATP7d和ATP1d上获得的aRMSE结果分别为0.045 0,0.012 8,0.006 5和0.006 8,都拥有非常小的aRMSE值,并且均比对比算法的aRMSE实验结果低,说明本文提出的LFS_MR算法具有很强的稳定性。

而本文的算法LFS_MR不仅进行了特征选择,还进行了样本选择,能够很好地去除冗余特征及噪音和离群点的干扰,通过低秩回归可使得多输出变量之间的关联被合理使用。四个数据集上的实验在aCC和aRMSE方面都取得了非常好的结果。因此,可证明本文提出的LFS_MR算法是一个能够很好地处理高维数据的多输出回归算法。

6 结束语

为能够很好处理高维数据的多输出回归问题,提出一种充分考虑特征与输出的关系,和输出变量之间的关系,以及样本之间的关系,使这三种关系得到很好结合的方法称为低秩特征选择多输出回归算法。为确保在学习模型的过程中利用多输出变量之间的关系来提高回归预测的准确度,该方法通过低秩约束去构建低秩回归模型;为能够去除数据中的噪音和离群点的干扰,利用有效的样本学习模型,该方法同时在该低秩回归上利用 $\ell_{2,p}-norm$ 进行样本选择;而且使用一个 $\ell_{2,p}-norm$ 正则化项惩罚回归系数矩阵,从特征与输出的关系方面进行特征选择去除冗余特征,解决高维数据的“维灾难”问题,可提高所获得的多输出回归模型的预测能力。通过实际数据集上实验的结果表明,LFS_MR算法处理高维数据的多输出回归分析能获得非常好的效果。

参考文献:

[1] Zhu Xiaofeng, Li Xuelong, Zhang Shichao.Block-row sparse multiview multilabel learning for image classification[J].IEEE Transactions on Cybernetics, 2016, 46(2): 450-461.

[2] Rai P, Kumar A, Daum´e III H.Simultaneously leveraging output and task structures for multiple-output regres-

sion[C]//Advances in Neural Information Processing Systems,2012:1-9.
[3] Anderson T.Estimating linear restrictions on regression coefficients for multivariate normal distributions[J].The Annals of Mathematical Statistics,1951,22(3):327-351.
[4] Argyriou A,Evgeniou T,Pontil M.Multi-task feature learning[C]//Advances in Neural Information Processing Systems,2006:41-48.
[5] Zhu Xiaofeng,Huang Zi,Shen Hengtao,et al.Dimensionality reduction by mixed kernel canonical correlation analysis[J].Pattern Recognition,2012,45(8):3003-3016.
[6] Donoho D.High-dimensional data analysis:the curses and blessings of dimensionality[C]//AMS Math Challenges Lecture,2000:1-32.
[7] Cai Xiao,Ding C,Nie Feiping.On the equivalent of low-rank regressions and linear discriminant analysis based regressions[C]//Proceedings of the 19th ACM SIGKDD,2013:1124-1132.
[8] Borchani H,Varando G,Bielza C,et al.A survey on multi-output regression[J].Data Mining and Knowledge,2015,5(5):216-233.
[9] Mukherjee A,Zhu Ji.Reduced rank ridge regression and its kernel extensions[J].Statistical Analysis and Data Mining,2011,4(6):612-622.
[10] Lu Canyi,Lin Zhouchen,Yan Shuicheng.Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization[J].IEEE Transactions on Image Processing,2015,24(2):646-654.
[11] Zhang Miao,Ding C,Zhang Ya,et al.Feature selection at the discrete limit[C]//Twenty-Eighth AAAI Conference on Artificial Intelligence,2014:1355-1361.
[12] Zhu Xiaofeng,Suk H I,Shen Dinggang.A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis[J].NeuroImage,2014,100:91-105.
[13] Zhang Shichao,Qin Zhenxing,Ling C X,et al."Missing is useful":missing values in cost-sensitive decision trees[J].IEEE Transactions on Knowledge and Data Engineering,2005,17(12):1689-1693.
[14] Qin Yongsong,Zhang Shichao,Zhu Xiaofeng,et al.Semi-parametric optimization for missing data imputation[J].Applied Intelligence,2007,27(1):79-88.
[15] Karalic A,Bratko I.First order regression[J].Machine Learning,1997,26(2):147-176.
[16] Spyromitros-Xioufis E,Groves W,Tsoumakas G,et al.Multi-label classification methods for multi-target regression[R].[S.l.]:Cornell University Library,2012:1159-1168.
[17] Wang Hua,Nie Feiping,Huang Heng,et al.Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance[C]//ICCV,2011:557-562.
[18] Cai Xiao,Nie Feiping,Cai Weidong,et al.New graph structured sparsity model for multi-label image annotations[C]//ICCV,2013:801-808.

(上接19页)

[8] Deng L,Yu D. Deep learning:methods and applications[J].Foundations and Trends in Signal Processing,2014,7:3-4.
[9] Schmidhuber J.Deep learning in neural networks:an overview[J].Neural Networks,2015,61:85-117.
[10] Bengio Y,LeCun Y,Hinton G. Deep learning[J].Nature,2015,521:436-444.
[11] Glauner P O.Deep convolutional neural networks for smile recognition[D].London:Imperial College,2015.
[12] Weng J.Brains as naturally emerging turing machines[C]//International Joint Conference on Neural Networks,2015:12-17.
[13] Gomes L.Machine-learning maestro Michael Jordan on the delusions of big data and other huge engineering efforts[J].IEEE Spectrum,2014.
[14] Lecun Y,Bengio Y,Hinton G E.Deep learning[J].Nature,2015,521(7553):436-444.
[15] Umamahesh S,Vishal M,Raghu R.SAR model[J].IEEE Transactions on Aerospace and Electronic Automatic Target Recognition using Discriminative Graphical System,2014,50(1):591-606.
[16] Fischer A,Igel C.Training restricted boltzmann machines:a introduction[J].Pattern Recognition,2014,47(1):25-39.
[17] Al-Masri E,Mahmoud Q H.Discovering the best web service[C]//16th International Conference on World Wide Web,2007:1257-1258.
[18] Al-Masri E,Mahmoud Q H.QoS-based discovery and ranking of web services[C]//IEEE 16th International Conference on Computer Communications and Networks,2007:529-534.
[19] Al-Masri E,Mahmooud Q H.Investigating web services on the world wide web[C]//17th International Conference on World Wide Web,2008:795-804.
[20] Web services:design,travel,shopping[EB/OL].[2017-02-20].http://www.ec-t.com/.
[21] 陈飞,刘奕群,张敏,等.基于查询子主题分类的多样性搜索评价方法[J].软件学报,2015,26(12):3130-3139.