基于超图表示的低秩属性选择方法用于回归分析*

钟 智[1]，方 月[2†]，胡荣耀[2]，李永钢[2]，雷 聪[2]
(1. 广西师范学院 计算机与信息工程学院，南宁 530023；2. 广西师范大学 广西多源信息挖掘与安全重点实验室，广西 桂林 541004)

摘 要： 针对现有的回归模型未考虑特征之间的深层结构，而导致在回归问题上输出不稳定的模型，提出了一种新的属性选择方法。通过稀疏学习理论中的 $\ell_{2,1}$-范数和 $\ell_{2,p}$-范数在线性回归模型分别进行样本降噪和属性选择，利用超图结构和低秩约束来分别考虑数据间的局部结构和不同数据间的全局结构，最后结合子空间学习方法来对模型进行微调。经实验证明，在回归分析中该算法较对比算法能取得更好的效果。

关键词： 低秩回归；属性选择；子空间学习；超图表示
中图分类号： TP391 **文献标志码：** A **文章编号：** 1001-3695(2018)07-2046-05
doi：10.3969/j.issn.1001-3695.2018.07.029

Based on hypergraph expressing low-rank feature selection algorithm for regression analysis

Zhong Zhi[1], Fang Yue[2†], Hu Rongyao[2], Li Yonggang[2], Lei Cong[2]
(1. *School of Computer & Information Engineering, Guangxi Teachers Education University, Nanning* 530023, *China*; 2. *Guangxi Key Laboratory of Multi-source Information Mining & Security, Guangxi Normal University, Guilin Guangxi* 541004, *China*)

Abstract: The traditional regression model does not consider the deep complex structure among the feature matrix so that it leads to output unstable regression model. To address this issue, this paper proposed a novel feature selection algorithm. Specifically, this paper first employed a $\ell_{2,1}$-norm regularizer and a $\ell_{2,p}$-norm regularizer, respectively, in the regression framework to conduct noisy reduction of samples and to select important features via considering the correlation among the features. Then it used a hypergraph representation and a low-rank constraint to consider the local structure and the global structure, respectively. Finally, it integrated subspace learning with the proposed model to further improve the result of feature selection. Experimental results show that the proposed method obtains the best results for regression analysis, compare to the four comparison methods.

Key words: low-rank regression; feature selection; subspace learning; hypergraph representation

在当今大数据环境下，各式各样的数据呈指数增长，尤其是在计算机视觉、生物医学等领域。高维数据不断地增加[1]，使得数据存储空间和数据处理的复杂度也随之增加。实际应用中，由于高维数据通常包含冗余属性而难以直接应用，所以利用属性约简来缩小高维数据的维度[2]，并筛选出最有意义的属性，以此来提高分类的准确率[3]。已有的属性约简方法可以分为子空间学习和属性选择两类[4,5]。子空间学习是将高维数据投影到低维空间，从而保持了数据间的相关结构。常见的方法有 fisher 线性鉴别分析(FLDA)[6]、局部判别式嵌入(LDE)法[7]、图嵌入(graph embedding)[8]等。属性选择方法是通过某种积分排名标准对属性进行排序，筛选出最有意义的特征，是一种提高算法性能的手段，因此它能够在模式识别和机器学习等领域[9,10]得到广泛应用。属性选择较为常见的方法分为有监督法、半监督法和无监督法三类。有监督的属性选择法有稀疏正则化线性回归[11]和稀疏逻辑回归法[12]，已经被证明在实际应用中经常超过无监督法和半监督法。因此，本文着重研究并提出新的有监督属性选择方法。

属性选择方法能够将高维数据进行降维，而子空间学习方法可以更好地保持数据间的相关结构。因此，本文在模型中同时运用属性选择和子空间学习法，提出一种基于超图表示的低秩属性选择方法用于回归分析(based on hypergraph expressing low-rank feature selection algorithm for regression analysis，LHSL_FS)。具体地，首先在线性回归模型框架中利用双重稀疏表示，即对损失项用 $\ell_{2,1}$-范数来有效地选取重要的样本，对正则化项用 $\ell_{2,p}$-范数进行属性选择。本文利用超图正则化项保持数据间的局部结构，用低秩约束来保持不同数据的全局结构；同时，在低秩模型中恰当地嵌入子空间学习法——LDA 算法对模型选择的结果进行微调；最后，探索出一种新的优化算法的方法，以此获得全局最优解。

1 相关工作

实际应用产生非常高维的数据，如信号处理(特别是音频和图像处理)、数据挖掘和多媒体数据检索、生物学和计算机图形学，并且需要选择一部分重要的特征(如信号源)而剔除那些噪声特征。属性选择可以用于解决这类型问题，即提取有意义的特征，丢弃噪声或冗余信息。例如，Du 等人[13]针对无标签数据，结合局部和全局图模型去探索数据的分布情况，利用局部判别模型和全局相关结构来提出有意义的特征。Cai 等人[14]使用 $\ell_{2,1}$-范数和 $\ell_{2,0}$-范数分别对损失函数和稀疏限制项进行约束来提取有用特征。Shi 等人[15]利用不同视角特征进行互补，并且利用 hessian 正则化对无标签进行编码，从而提高特征提取的效率。但是属性选择直接摈弃一部分特征会丢失一些重要信息。子空间学习主要应用在由于数据维度高导致分类模型过拟合的场景。高维数据实际上就是子空间中的无

收稿日期：2017-02-25；修回日期：2017-04-10 基金项目：国家自然科学基金资助项目(61450001，61263035，61573270)；国家"973"计划资助项目(2013CB329404)；中国博士后科学基金资助项目(2015M570837)；广西自然科学基金资助项目(2015GXNSFCB139011，2015GXNSFAA139306)；广西研究生教育创新计划资助项目(XYCSZ2017064)

作者简介：钟智(1963-)，男，广西梧州人，副教授，主要研究方向为机器学习、数据挖掘；方月(1992-)，男(通信作者)，湖北孝感人，硕士，主要研究方向为数据挖掘、机器学习(854289665@qq.com)；胡荣耀(1992-)，男，江西景德镇人，硕士，主要研究方向为数据挖掘、机器学习；李永钢(1989-)，男，河北保定人，硕士，主要研究方向为数据挖掘、机器学习；雷聪(1991-)，男，湖北大冶人，硕士，主要研究方向为数据挖掘、机器学习.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

噪声数据和高维噪声数据的叠加,而这样的子空间维度较低,如果能找出这样的子空间,就可以保证在不丢失重要特征的前提下实现数据降维。例如,Sui 等人[16]利用暂时获得的目标的局部相互关系,将逐行稀疏引入到目标子空间,因此能够适应性地忽略干扰信息。Peng 等人[17]基于 ℓ_2-范数建立图,称做 L_2-graph 方法,并用于子空间聚类和子空间学习,它可以探索数据之间的隐藏结构。Li 等人[18]通过将图像识别和特征学习整合到联合学习框架中,提出了一种新颖的鲁棒结构化子空间学习算法(RSSL),将学习到的子空间看做媒介空间,以减少低级视觉特征与高级语义之间的语义差距。但是单纯地使用子空间学习会丢失数据之间更深层次的结构。

由于现实问题中数据之间的关系错综复杂,仅仅用普通图来表示数据间的相关性会导致信息的大量流失。然而一种补救大量信息损失的方式就是使用超图代替普通图来表示数据间的相关性。例如,Zhang 等人[19]提出了一种基于超图的信息理论属性选择算法,首先让数据的每个节点对应于其特征去构建超图,每条边有一个相关权重对应于该边连接的多维属性;然后应用超图到超图聚类以定位最有益的特征子集,因此特征子集具有较低的冗余度和很强的鉴别能力。

2　相关理论背景及简介

2.1　稀疏学习简介

稀疏学习(sparse learning)[20]近年来在机器学习、模式识别等领域[21]成为应用热点。由于属性选择是考虑属性的稀疏性,通过属性选择过滤冗余属性,使得学习器训练过程在低维空间中进行。设 $Z=[z_1,z_2,\cdots,z_n]\in \mathrm{R}^{n\times c}$ 是模型重构系数矩阵,$z_1,z_2,\cdots,z_n$ 表示数据点之间的重构系数。属性选择方法利用稀疏编码理论,采用范数正则化惩罚目标函数,使系数矩阵变得稀疏。该类问题可概括为求解如下模型:

$$\min_Z g(Z),g(Z)=\mu(Z)+\lambda\varphi(Z) \tag{1}$$

其中:$\mu(Z)$ 是损失函数;$\varphi(Z)$ 是正则化项;λ 为控制模型系数矩阵 Z 的稀疏性。若重构系数为 0,则对应样本不进行重构,并且选择出系数矩阵中非零元素对应的样本,保留了样本的相关性。属性选择算法使用稀疏学习能够保留原始数据的系数权重,通过稀疏约束去除无关属性。通常稀疏学习选用能够凸优化求解范数作为正则化因子,并以此来使得学习到的模型能够获得全局最优解[22]。本文利用 $\ell_{2,1}$、$\ell_{2,p}$ 凸优化求解范数,$\ell_{2,1}$-范数能导致行稀疏,可以去除样本间的差异,而 $\ell_{2,p}$-范数可通过参数 p 来控制属性间关联结构的程度,能够使系数矩阵更稀疏,以达到剔除冗余和不相关属性的目的。

2.2　超图简介

定义三元组 $G_H=(V,E,w)$ 表示超图,其中:$V=\{1,\cdots,n\}$ 表示顶点;E 是 V 的非空子集表示超边;w 是超边权重矩阵。本文把 $w(ei)$ 设置为 0 或 1。图 1(a)中,$E=\{e_1=\{v_1,v_2,v_3\},e_2=\{v_3,v_4,v_5\},e_3=\{v_5,v_6,v_7\},e_4=\{v_1,v_7,v_8\}\}$ 表示超边的集合。超边 e_1 是由顶点 v_1、v_2、v_3 组成;超边 e_2 是由顶点 v_3、v_4、v_5 组成;超边 e_3 是由顶点 v_5、v_6、v_7 组成;超边 e_4 是由顶点 v_1、v_7、v_8 组成。图 1(b)表示图 1(a)中顶点与超边具体数值的关系。超图 G_H 是描述顶点与边的关系矩阵,即用 $H\in \mathrm{R}^{|V|\times|E|}$ 表示关联矩阵,定义如式(2)所示。

$$H(v,e)=\begin{cases}1 & \text{if}\quad v\in e\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

根据超图 G_H 描述的关系矩阵,顶点 v 的度 $d(v)$ 和超边 e 的度 $\delta(e)$ 表示为

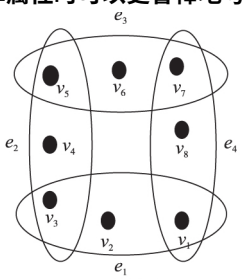
$$d(v)=\sum_{e\in E}w(e)H(v,e),\delta(e)=\sum_{v\in V}H(v,e) \tag{3}$$

实际中通常采用规范化的拉普拉斯方法去建立超图拉普拉斯矩阵[23],定义如式(4)所示。

$$L_H=I-D_V^{-\frac{1}{2}}HW_HD_e^{-1}J^TD_v^{-\frac{1}{2}} \tag{4}$$

其中:I 是一个单位矩阵;D_v 表示顶点的对角矩阵,其对角元素 $d(v)$ 是关联矩阵 H 第 i 行之和;W_H 表示由超边权重矩阵 w 构成的对角矩阵;D_e 表示超边的对角矩阵,其对角元素由 $\delta(e)$ 构成,表示关联矩阵 H 第 j 列之和。

超图描述了超边与多顶点的关系,超边可以包含许多样本点的子集,超边中的样本数代表度。利用超图表示使得模型在选择属性时可以更鲁棒地考虑数据之间的局部结构。



(a)超图样例

	e_1	e_2	e_3	e_4
v_1	1	0	0	1
v_2	1	0	0	0
v_3	1	1	0	0
v_4	0	1	0	0
v_5	0	1	1	0
v_6	0	0	1	0
v_7	0	0	1	1
v_8	0	0	0	1

(B)顶点与边的关联矩阵

图1　超图与关系矩阵表

3　算法描述和优化

3.1　算法描述

基本线性模型通过属性的线性组合来进行预测。具体模型为

$$\min_Z \|Y-X^TZ\|_F^2 \tag{5}$$

其中:$\|\cdot\|_F$ 是 frobenius 范数,定义为 $\|S\|_F=(\sum_{i=1}^{n}\sum_{j=1}^{m}|s_{ij}|^2)^{\frac{1}{2}}$;$X=[x_1,x_2,\cdots,x_n]\in \mathrm{R}^{d\times n}$ 表示训练样本矩阵,n 和 d 分别表示样本个数和属性个数;$Y\in \mathrm{R}^{n\times c}$ 表示类标签矩阵,n 和 c 分别表示输入数据的样本数和类别数。模型输出的系数矩阵 $Z\in \mathrm{R}^{d\times c}$ 代表训练样本与测试样本间重建的系数矩阵。考虑到样本间的差异可能比较大,而 $\ell_{2,1}$-范数可以导致行稀疏,比 ℓ_F-范数能更好地去除样本间差异的影响,因此得到的改进模型为

$$\min_Z \|Y-X^TZ\|_{2,1} \tag{6}$$

其中:$\|\cdot\|_{2,1}$ 是 $\ell_{2,1}$ 矩阵范数,定义为 $\|S\|_{2,1}=\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}|s_{ij}|^2}$。但是该方法没有考虑输出之间的关系,在多分类应用中输出之间的隐含关系会更加明显[22,24]。

为了让式(6)学习得到的模型更具有泛化能力,常用的做法是保持样本的局部结构。例如局部保持投影(LPP)[25]:首先构建一个包含数据集邻域信息的图,然后使用图拉普拉斯计算一个将数据点映射到子空间的变换矩阵。这种线性变换在某种意义上最佳地保留了数据的局部邻域信息,LPP 方法的局部保留特性使得它对异常值和噪声相对不敏感,但是 LPP 构建的图不能表示复杂的数据间关系。2.2 节显示超图可以解决这个问题,即学习到数据间复杂的局部相关结构。根据 2.2 节所描述,本文将超图嵌入到模型中,嵌入的超图拉普拉斯正则化项为 $R_1(Z)=\alpha\mathrm{tr}(Z^TXL_HX^TZ)$,这样使得通过空间投影变换后的样本数据的近邻关系结构不变。所以式(6)可以转换成如下问题:

$$\min_Z \|Y-X^TZ\|_{2,1}+\alpha\mathrm{tr}(Z^TXL_HX^TZ) \tag{7}$$

其中:L_H 为通过式(4)得到的超图 Laplacian 矩阵;α 为正则化控制参数。给定样本数据 X,首先得到顶点与边的关联矩阵,然后通过规范化的拉普拉斯方法建立超图 Laplacian 矩阵 L_H。此时,式(7)中用超图代替普通图,避免了因使用普通图而导致不能发现数据间的复杂关系这一问题。为了去除冗余属性和离群数据,并且让学习得到的模型在进行属性选择时能够有效地避免维度灾难问题[26],在模型中引入 $\ell_{2,p}$-范数正则化项,即 $R_2(Z)=\|Z\|_{2,p}=(\sum_{i=1}^{d}(\sum_{j=1}^{c}Z_{ij}^2)^{p/2})^{1/p}$,对回归系数矩阵 Z 进行惩罚。加入惩罚项后的新模型如下:

$$\min_Z \|Y-X^TZ\|_{2,1}+\alpha\mathrm{tr}(Z^TXL_HX^TZ)+\beta\|Z\|_{2,p} \tag{8}$$

其中:β 为调节行稀疏正则化控制参数。为了充分考虑并利用各输出之间的关系,本文对系数矩阵 Z 进行低秩结构约束[27],即令 $Z=AB$ 且 $\mathrm{rank}(Z)=r,r<\min(n,c)$。这样既考虑到属性

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

间的关联结构，又能提高算法运行效率。最终得到的模型如下：

$$\min_{A,B} \| Y - X^T AB \|_{2,1} + \alpha \mathrm{tr}(B^T A^T X L_H X^T AB) + \beta \| AB \|_{2,p} \text{ s.t. } \mathrm{rank}(AB) \leqslant \min(n, c) \tag{9}$$

其中：AB 是由 $A \in \mathrm{R}^{d \times r}$ 和 $B \in \mathrm{R}^{r \times c}$ 构成的系数矩阵。属性间隐藏的关系可以用低秩限制来探索。使用低秩限制实际上就是使用 LDA 进行子空间学习[28]，而子空间学习尽量保证不丢失信息并且能降低数据的维度。通过稀疏表示，使得系数矩阵 AB 中大部分行元素变为 0，AB 中非零元素位置对应的属性得到保留，因此达到属性选择的目的。而且超图可以保留各数据之间的复杂局部相关性结构，即通过每个样本的最近邻表示一组数据集的局部邻域关系。这种真正探索数据间复杂结构的方法，显然能使得模型更具有稳定性。

本文提出的 LHSL_FS 算法用于回归分析具有以下优点：

a) LHSL_FS 算法在属性选择时考虑了类标签之间的关系，并且运用了双重稀疏模式，即对损失项用 $\ell_{2,1}$-范数进行样本选择，对回归系数矩阵 AB 用 $\ell_{2,p}$-范数进行惩罚，通过调节 p $(0 < p < 2)$ 使得系数矩阵更稀疏，从而使模型能准确地筛选重要的属性。因此，该算法具有较好的鲁棒性和泛化性。

b) 为了合理考虑各类数据间的关联结构，用超图正则化项保持数据间的局部结构，用低秩约束来保持不同数据的全局结构。同时为了让模型选择出来的属性更具有代表性，在模型中恰当地嵌入经典子空间学习法——LDA 算法，以此用于对低秩属性模型选择的结果进行微调。因此，该算法可以更好地处理高维数据的多输出回归分析问题。

c) 对本文的目标函数提出一种区别于交替方向乘子法的求解方法，首先固定低秩属性选择的结果，以此增强子空间学习的能力；然后固定子空间学习的结果，确保低秩属性选择能输出更具判别力的属性集。该优化算法使目标值在每次迭代过程中逐步趋近于全局最优解，最终取得全局最优解。

本文提出的算法具体实现步骤如算法 1 所示。

算法 1　LHSL_FS 算法伪代码

输入：训练样本 $X \in \mathrm{R}^{d \times n}$，类标签 $Y \in \mathrm{R}^{n \times c}$，控制参数 α、β、p。

输出：aCC 和 aRMSE。

1 通过训练样本得出类指示矩阵 $Y = [y_{i,c}]$，初始化关系矩阵 H，通过式(4)计算超图拉普拉斯矩阵 L_H；

2 根据式(9)调用算法2 求解全局最优解，得到系数矩阵 $Z^* = A^* B^*$；

3 通过最优解 Z^* 对数据集 X 的属性进行筛选得到新属性集；

4 对新属性集构成的样本采用 SVM 分类，求出测试样本的预测值；

5 计算 aCC 和 aRMSE 的对应结果。

3.2　算法优化

首先对式(9)进行化简得到

$$\min_{A,B} \mathrm{tr}((Y - X^T AB)^T D (Y - X^T AB)) + \alpha \mathrm{tr}(B^T A^T X L_H X^T AB) + \beta \mathrm{tr}(B^T A^T QAB) \tag{10}$$

其中：$D \in \mathrm{R}^{n \times n}$、$Q \in \mathrm{R}^{d \times d}$ 分别表示 $Y - X^T AB$ 和 AB 的对角矩阵。D 中的对角线上的每个元素取值如下：

$$d_{ii} = \frac{1}{2 \| u^i \|_2^1} \quad \text{s.t.} \quad i = 1, 2, \cdots, n \tag{11}$$

其中：u^i 表示矩阵 $U^* = (Y - X^T A^* B^*)$ 的第 i 行；A^* 和 B^* 分别表示矩阵 A 和 B 的最优解。Q 中的对角线上的每个元素取值如下：

$$q_{jj} = \frac{1}{(2/p)(\| g^j \|_2)^{2-p}} \quad \text{s.t.} \quad j = 1, 2, \cdots, d; 0 < p < 2 \tag{12}$$

其中：g^j 表示矩阵 $G^* = A^* B^*$ 的第 j 行，而 G^* 为重构系数矩阵 Z 的最优解。令 $J(A,B)$ 表示式(10)，并对 B 求偏导数，然后令求到的偏导数等于零。

$$\frac{\partial J(A,B)}{\partial B} = -2A^T XDY + 2A^T XDX^T AB + 2\alpha A^T X L_H X^T AB + 2\beta A^T QAB = 0 \tag{13}$$

式(13)中本文用 A 表示 B 得到

$$B = (A^T (XDX^T + \alpha X L_H X^T + \beta Q) A)^{-1} A^T XDY \tag{14}$$

然后将矩阵 B 代入式(10)化简得到

$$\max_A \mathrm{tr}((A^T (XDX^T + \alpha X L_H X^T + \beta Q) A)^{-1} A^T XDYY^T D^T X^T A) \tag{15}$$

注意到：

$$S_t = XDX^T + \alpha X L_H X^T + \beta Q, S_b = XDYY^T D^T X^T \tag{16}$$

其中：S_t 和 S_b 分别是 LDA 中类内离散矩阵和类间离散度矩阵。式(15)只含有一个未知量 A，因此本文求解得到 A 为

$$A = \arg\max_A \{ \mathrm{tr}(A^T S_t A)^{-1} A^T S_b A) \} \tag{17}$$

式(17)是 LDA 求解问题，因此，上式对应的全局最优解为：求 $S_t^{-1} S_b$ 中非零属性值所对应的属性向量。

优化求解式(9)具体步骤如算法 2 所示。

算法 2　优化求解式(9)的伪代码

输入：训练样本 $X \in \mathrm{R}^{d \times n}$，类标签 $Y \in \mathrm{R}^{n \times c}$，控制参数 α、β、r、p。

输出：矩阵 A 和 B。

1 初始化 $t = 1$；

2 随机初始化矩阵 $A^{(t)}$ 和 $B^{(t)}$，计算 $Z^{(t)} = A^{(t)} B^{(t)}$；

3 初始化 $D^{(t)} = I \in \mathrm{R}^{n \times n}$，$Q^{(t)} = I \in \mathrm{R}^{d \times d}$；

4 重复：

5 通过式(17)和(14)分别计算 $A^{(t+1)}$ 和 $B^{(t+1)}$；

6 由式(11)(12)分别更新 $D^{(t+1)}$ 和 $Q^{(t+1)}$；

7 更新 $t = t + 1$；

8 直到式(9)收敛；

9 结束。

4　优化算法收敛性证明

式(9)的目标函数是凸的但是非光滑的，并且有两个未知变量 A、B 需要被优化，导致目标函数难以直接求解。本文提出一种简单有效的算法来求解该问题，其具体收敛性证明步骤如下：算法 2 的步骤 4 ~ 9 称为一次迭代过程，假如迭代到第 t 次，则有

$$\langle A^{(t+1)}, B^{(t+1)} \rangle = \arg\min_{A,B} \| (Y - X^T A^{(t)} B^{(t)}) \|_{2,1} + \alpha \mathrm{tr}(B^{(t)T} A^{(t)T} X L_H X^T A^{(t)} B^{(t)}) + \beta \mathrm{tr}(B^{(t)T} A^{(t)T} Q^{(t)} A^{(t)} B^{(t)}) \tag{18}$$

因此可以得到

$$\| Y - X^T A^{(t+1)} B^{(t+1)} \|_{2,1} + \alpha \mathrm{tr}(B^{(t+1)T} A^{(t+1)T} X L_H X^T A^{(t+1)} B^{(t+1)}) + \beta \mathrm{tr}(B^{(t+1)T} A^{(t+1)T} Q^{(t)} A^{(t+1)} B^{(t+1)}) \leqslant \| Y - X^T A^{(t)} B^{(t)} \|_{2,1} + \alpha \mathrm{tr}(B^{(t)T} A^{(t)T} X L_H X^T A^{(t)} B^{(t)}) + \beta \mathrm{tr}(B^{(t)T} A^{(t)T} Q^{(t)} A^{(t)} B^{(t)}) \tag{19}$$

令 $G^{(t)} = A^{(t)} B^{(t)}$，$G^{(t+1)} = A^{(t+1)} B^{(t+1)}$，把式(12)的对角矩阵 Q 代入式(19)中可以得到

$$\| (Y - X^T G^{(t+1)}) \|_{2,1} + \alpha \mathrm{tr}(G^{(t+1)T} X L_H X^T G^{(t+1)}) + \beta \sum_{i=1}^{d} \frac{(\| g^{i(t+1)} \|_2)^{2(2-p)}}{(2/p)(\| g^{i(t)} \|_2)^{(2-p)}} \leqslant \| (Y - X^T G^{(t)}) \|_{2,1} + \alpha \mathrm{tr}(G^{(t)T} X L_H X^T G^{(t)}) + \beta \sum_{i=1}^{d} \frac{(\| g^{i(t)} \|_2)^{2(2-p)}}{(2/p)(\| g^{i(t)} \|_2)^{(2-p)}} \tag{20}$$

其中：$g^{i(t)}$ 和 $g^{i(t+1)}$ 分别是 $G^{(t)}$ 和 $G^{(t+1)}$ 的第 i 行。而对于每个 i，可以得到以下不等式：

$$\| (g^{i(t+1)} \|_2)^{(2-p)} - \frac{(\| g^{i(t+1)} \|_2)^{2(2-p)}}{(2/p)(\| g^{i(t)} \|_2)^{(2-p)}} \leqslant (\| g^{i(t)} \|_2)^{(2-p)} - \frac{(\| g^{i(t)} \|_2)^{2(2-p)}}{(2/p)(\| g^{i(t)} \|_2)^{(2-p)}} \tag{21}$$

然后，将上面的不等式乘以正则化参数 β 并进行累加后，可以得到

$$\beta \sum_{i=1}^{d} (\| g^{i(t+1)} \|_2)^{(2-p)} - \frac{(\| g^{i(t+1)} \|_2)^{2(2-p)}}{(2/p)(\| g^{i(t)} \|_2)^{(2-p)}} \leqslant \beta \sum_{i=1}^{d} (\| g^{i(t)} \|_2)^{(2-p)} - \frac{(\| g^{i(t)} \|_2)^{2(2-p)}}{(2/p)(\| g^{i(t)} \|_2)^{(2-p)}} \tag{22}$$

结合式(20)和(22)就可以得到

$$\| (Y - X^T G^{(t+1)}) \|_{2,1} + \alpha \mathrm{tr}(G^{(t+1)T} X L_H X^T G^{(t+1)}) + \beta \sum_{i=1}^{d} (\| g^{i(t+1)} \|_2)^{2(2-p)} \leqslant \| (Y - X^T G^{(t)}) \|_{2,1} + \alpha \mathrm{tr}(G^{(t)T} X L_H X^T G^{(t)}) + \beta \sum_{i=1}^{d} (\| g^{i(t)} \|_2)^{2(2-p)} \tag{23}$$

综上所述，可得到如下结果：

$$\| (Y - X^T G^{(t+1)}) \|_{2,1} + \alpha \mathrm{tr}(G^{(t+1)T} X L_H X^T G^{(t+1)}) + \beta \| G^{(t+1)} \|_{2,p} \leqslant$$

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\|(Y-X^TG^{(t)})\|_{2,1}+\alpha\mathrm{tr}(G^{(t)T}XL_HX^TG^{(t)})+\beta\|G^{(t)}\|_{2,p} \quad (24)$$

不等式(24)说明算法 1 在每次迭代中目标函数都是单调递减的,因此算法 1 可以达到全局最优解。

5　实验结果和分析

5.1　实验数据集和对比算法

多目标回归(MTR)也称为多变量或多输出回归,是指使用一组常用的输入变量来预测多个连续变量的任务[29]。本文使用两种公开数据集(EDM、SF1)和四种基准数据集(ATP1d、ATP7d、OES97、OES10),这六种数据集都是文本数据,可用于多目标回归(MTR)。

EDM(electrical discharge machining)[30]数据集用于二类目标回归问题。任务是通过控制两个变量值再现操作人员的行为来缩短机器加工时间。每个目标变量可取 3 个不同的数值{-1,0,1},有 16 个连续的输入变量。此数据集具有 154 个样本,每个样本 16 个属性,每个属性取值为[-4.96,6.1],2 种输出目标,可用于二类回归问题。

SF(solar flare)[31]数据集有 3 个目标变量对应于三种太阳斑点(普通、中等、严重),是在 24 h 内观察到的数据。有 2 个版本,分别是 SF_1 和 SF_2。SF_1 包含 1969 年的观测数据,具有 323 个样本,每个样本 10 个属性,每个属性可取值{0,1,2,3,4},3 种输出目标,每个目标值取值在[1,24],可用于多类回归问题。

ATP(airline ticket price)[29]数据集的 6 个目标变量是第二天价格(ATP1d)或在接下来的 7 d 观察的最低价格(ATP7d)。ATP1d 和 ATP7d 数据集分别具有 337 和 296 个样本,每个样本都具有由 411 个属性,每个属性取值在[0,3573],6 种输出目标,可用于多类回归问题。

OES(occupational employment survey)[29]数据集是由美国劳工统计局从 1997 年来调查得到的职业就业调查数据(OES97)和 2010(OES10)的年度就业调查数据,分别有 334 和 403 个城市的就业调查数据。OES97 数据集具有 334 个样本,每个样本具有由 263 个属性,每个属性取值为[30,126190],16 种输出目标,可用于多类回归问题。OES10 数据集具有 403 个样本,每个样本具有 298 个属性,每个属性取值为[30,156740],16 种输出目标,可用于多类回归问题。

数据集信息统计如表 1 所示。表 1 中第一列是数据集名称,第二列是样本数量,第三列是属性数目,第四列是输出类别数。在表 1 中本文列出了数据集的详细分布结构。

表 1　数据集信息统计

数据集	样本数	属性数	多输出类别	数据集	样本数	属性数	多输出类别
EDM	154	16	2	ATP7d	296	411	6
SF1	323	10	3	OES97	334	263	16
ATP1d	337	411	6	OES10	403	298	16

实验在 Windows 7 系统上运行,用 MATLAB 语言编写代码,并选用四种对比算法与 LHSL_FS 进行比较。对比算法介绍如下:

a) LSG21 方法[32]。它是一种新的结构稀疏诱导范数正则化算法,它将关系图信息融合到多标签分类模型中,在系数矩阵上应用 $\ell_{2,1}$-范数去寻找与输入相关的共同输出。相比 LHSL_FS 算法,LSG21 方法没有使用低秩约束去除冗余属性或离群数据点。

b) CSFS 方法[33]。它是一种凸的半监督多标签属性选择算法,相对于传统的基于图的半监督属性选择算法,它不需要构造图,并且应用 $\ell_{2,1}$-范数进行稀疏属性选择。对比 LHSL_FS 算法,由于没有使用低秩约束,忽略了数据之间的全局结构,所以在回归模型上表现不理想。

c) SLRR 方法[24]。它是一种稀疏低秩回归模型,利用低秩约束寻求数据中的低秩结构和应用系数矩阵范数进行特征选择。相比 LHSL_FS 方法而言,在回归模型中无超图拉普拉斯正则化项,因此不能保持数据之间的局部结构。

d) RSR 方法[34]。它通过自表征的方法选择一个最具代表性的响应矩阵,然后嵌入到稀疏学习模型中进行属性选择。虽然对数据进行了相应的稀疏处理,但是相比 LHSL_FS 方法而言,缺乏类标签的有效辅助。

本文用平均相关系数(average correlation coefficient,aCC)和均方差误差(average root mean square error,aRMSE)[35]作为评价指标。其中,均方差误差可以反映算法的稳定性,均方差值越小,代表算法的稳定性越好;而平均相关系数表示预测值与测试样本目标值之间的相关性,其值越大,说明预测值越接近测试目标值。

5.2　实验结果和分析

对比算法和 LHSL_FS 算法的实验均采用 10-折交叉验证法,并且在 10-折交叉验证的每一折中再进行 5-折交叉验证,用 SVR 训练来学习回归模型。在模型的参数设置中,本文把控制参数 α 和 β 设置为 $(\alpha,\beta)\in[10^{-3},10^{-2},10^{-1},1,10^{1},10^{2},10^{3}]$,低秩参数 r 设置为 $r<\min(n,c)$。

六种数据集分别如图 2 ~ 7 所示。通过图 2 ~ 7 可以清楚地看出,LHSL_FS 算法在六个数据集上的 aCC,由于 10 折交叉验证的随机性,并不能保证每次的实验结果都是最好的,但是在每个数据集上 10 次实验结果比对比算法大部分都较高,最终的 aCC 也是最高的。

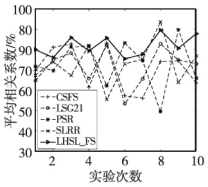


图2　数据集EDM

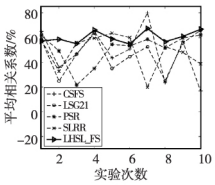


图3　数据集SF1

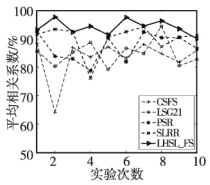


图4　数据集ATP1d

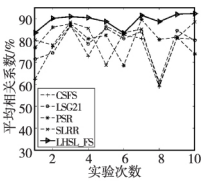


图5　数据集ATP7d

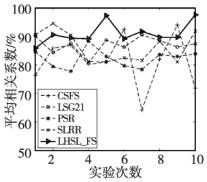


图6　数据集OES97

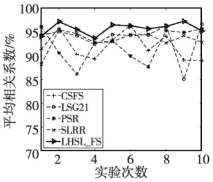


图7　数据集OES10

aCC 统计结果如表 2 所示。从表 2 数据可以看出,LHSL_FS 算法与其他四种对比算法在所有数据集上比较,均获得了最高的 aCC。具体地,LHSL_FS 算法在 OES10 数据集上取得了最优的结果,其 aCC 为95.13%,相比 LSG21、CSFS、SLRR 和 RSR 算法分别高出 0.63%、2.96%、0.98%、2.45%。另外,在 ATP1d 数据集上,LHSL_FS 算法提高了 9.25%、10.71%、4.09%、7.08%。LHSL_FS 算法在 SF1 数据集上的 aCC 只有 56.27%,但相比对比算法也分别高出 5.52%、3.32%、1.35%、2.70%。在所有数据集上,相对于对比算法,LHSL_FS 算法的 aCC 分别提高了 6.06%、7.86%、4.26%、5.75%。

表 2　aCC 统计结果　/%

数据集	算法				
	LSG21	CSFS	SLRR	RSR	LHSL_FS
EDM	72.22	71.05	73.44	73.75	80.83
SF1	50.75	52.95	54.92	53.57	56.27
ATP1d	84.89	83.43	90.05	87.06	94.14
ATP7d	78.74	77.00	79.91	80.77	89.15
OES97	87.85	81.57	87.29	82.98	89.80
OES10	94.50	92.17	94.15	92.68	95.13
average	78.16	76.36	79.69	78.47	84.22

四种对比算法在这六个不同数据集上都获得了很好的 aCC 结果。但是 LSG21 算法没有对样本进行选择,并且在构建图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。CSFS 算法通过应用 $\ell_{2,1}$-范数约束来进行属性选择,但是忽略了数据之间的关联结构,在回归模型上获得的 aCC 结果不是很理想。SLRR 算法没有对样本进行选择,数据中的噪声和离群点会干扰回归模型的学习,其获得的 aCC 结果比 LHSL_FS 的差。RSR 算法虽然对数据进行了相应的稀疏处理,但是没有

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

考虑类标签与数据之间的相关性。

另外，LSG21、CSFS、SLRR、RSR 和本文提出的 LHSL_FS 算法在上述六个数据集上实验的 aRMSE 结果如表 3 所示。

表 3 所有对比算法在六个数据集上的 aRMSE 结果

数据集	算法				
	LSG21	CSFS	SLRR	RSR	LHSL_FS
EDM	0.044 4	0.045 5	0.046 0	0.046 1	0.044 0
SF1	0.050 5	0.042 7	0.041 4	0.021 9	0.019 9
ATP1d	0.015 5	0.015 3	0.008 4	0.009 3	0.006 6
ATP7d	0.014 7	0.015 3	0.009 0	0.008 8	0.006 6
OES97	0.015 7	0.017 4	0.014 8	0.016 3	0.014 6
OES10	0.012 7	0.014 4	0.0132	0.012 6	0.012 6
average	0.025 6	0.025 1	0.022 1	0.019 2	0.017 4

从表 3 可知，LHSL_FS 算法在数据集 EDM、SF1、ATP1d、ATP7d、OES97 和 OES10 上获得的 aRMSE 结果分别为 0.044 0、0.019 9、0.006 6、0.006 6、0.014 6 和 0.012 6，都拥有非常小的 aRMSE 值，并且均比对比算法的 aRMSE 实验结果低。LSG21、CSFS、SLRR、RSR 和 LHSL_FS 算法在六个数据集上的平均 aRMSE 值分别为 0.025 6、0.025 1、0.022 1、0.019 2、0.017 4，可以看出 LHSL_FS 算法比对比算法在平均 aRMSE 上都低，说明本文提出的 LHSL_FS 算法具有比较强的稳定性。由于不同数据集有不同的数据分布，且含有不同的干扰因素，本文提出的 LHSL_FS 算法用于回归分析在每个数据集上都要优于对比算法，相比对比算法而言，LHSL_FS 算法具有最高的鲁棒性，同时也说明 LHSL_FS 算法输出了更具判别力的属性集。同时，与子空间学习方法相比，本文算法不仅保留了提取后的重要属性，还利用子空间学习对重要属性进一步微调，相比单一的属性选择算法保证了数据自身的全局结构，所以使得本文算法具有更好的解释性。

6 结束语

本文提出了一种综合性的属性选择算法（LHSL_FS 算法）用于回归分析。算法有效地在一个经典的回归模型中融合了子空间学习和低秩超图属性选择，既扩展了子空间学习在回归模型上的应用，也弥补了低秩属性选择在保持数据几何结构方面的不足。实验表明 LHSL_FS 算法在处理高维数据的多输出回归分析方面能获得很好的结果。在今后的研究中，将尝试使用半监督属性选择方法扩展本文提出的算法，并尝试使用自表征等技术来改进算法。

参考文献:

[1] Zhang Shichao, Jin Zhi, Zhu Xiaofeng. Missing data imputation by utilizing information within incomplete instances [J]. Journal of Systems and Software, 2011, 84(3): 452-459.

[2] Zhu Xiaofeng, Huang Zi, Shen Hengtao, *et al*. Dimensionality reduction by mixed kernel canonical correlation analysis [J]. Pattern Recognition, 2012, 45(8): 3003-3016.

[3] Zhu Xiaofeng, Zhang Shichao, Jin Zhi, *et al*. Missing value estimation for mixed-attribute data sets [J]. IEEE Trans on Knowledge and Data Engineering, 2011, 23(1): 110-121.

[4] Zhao Yanchang, Zhang Shichao. Generalized dimension-reduction framework for recent-biased time series analysis [J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(2): 231-244.

[5] Zhu Xiaofeng, Suk H I, Shen Dinggang. Matrix-similarity based loss function and feature selection for alzheimer's disease diagnosis [C]//Proc of Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 3089-3096.

[6] Welling M. Fisher linear discriminant analysis [J]. Department of Computer Science, 2007, 16(94): 237-280.

[7] Chen H T, Chang Huangwei, Liu T L. Local discriminant embedding and its variants [C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2005: 846-853.

[8] Yan Shicheng, Xu Dong, Zhang Benyu, *et al*. Graph embedding: a general framework for dimensionality reduction [C]//Proc of Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2005: 830-837.

[9] Zhu Xiaofeng, Li Xuelong, Zhang Shichao. Block-row sparse multiview multilabel learning for image classification [J]. IEEE Trans on Cybernetics, 2015, 46(2): 450-461.

[10] Qin Yongsong, Zhang Shichao, Zhu Xiaofeng, *et al*. Semi-parametric optimization for missing data imputation [J]. Applied Intelligence, 2007, 27(1): 79-88.

[11] Zhang Shichao, Li Xuelong, Zong Ming, *et al*. Learning k for KNN classification [J]. ACM Trans on Intelligent Systems and Technology, 2017, 8(3): 43.

[12] Liimatainen K, Heikkilä R, Yli-Harja O, *et al*. Sparse logistic regression and polynomial modelling for detection of artificial drainage networks [J]. Remote Sensing Letters, 2015, 6(4): 311-320.

[13] Du Xingzhong, Yan Yan, Pan Pingbo, *et al*. Multiple graph unsupervised feature selection [J]. Signal Processing, 2016, 120(3): 754-760.

[14] Cai Xiao, Nie Feifeng, Huang Heng. Exact top-k feature selection via $\ell_{2,0}$-norm constraint [C]//Proc of the 23rd International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2013: 1240-1246.

[15] Shi Caijuan, An Gaoyun, Zhao Ruizhen, *et al*. Multi-view Hessian semi-supervised sparse feature selection for multimedia analysis [J]. IEEE Trans on Circuits & Systems for Video Technology, 2017, 27(9): 1947-1961.

[16] Sui Yao, Zhang Shunli, Zhang Li. Robust visual tracking via sparsity induced subspace learning [J]. IEEE Trans on Image Processing, 2015, 24(12): 4686-4700.

[17] Peng Xi, Yu Zhiding, Yi Zhang, *et al*. Constructing the L2-graph for robust subspace learning and subspace clustering [J]. IEEE Trans on Cybernetics, 2017, 47(4): 1053-1066.

[18] Li Zechao, Liu Jing, Tang Jinhui, *et al*. Robust structured subspace learning for data representation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2015, 37(10): 2085-2098.

[19] Zhang Zhihong, Hancock E R. Hypergraph based information-theoretic feature selection [J]. Pattern Recognition Letters, 2012, 33(15): 1991-1999.

[20] Zhu Xiaofeng, Huang Zi, Yang Yang, *et al*. Self-taught dimensionality reduction on the high-dimensional small-sized data [J]. Pattern Recognition, 2013, 46(1): 215-229.

[21] Zhu Xiaofeng, Li Xuelong, Zhang Shichao, *et al*. Robust joint graph sparse coding for unsupervised spectral feature selection [J]. IEEE Trans on Neural Networks and Learning Systems, 2017, 28(6): 1263-1275.

[22] Zhu Xiaofeng, Huang Zi, Cheng Hong, *et al*. Sparse hashing for fast multimedia search [J]. ACM Trans on Information Systems, 2013, 31(2): 9.

[23] Li Yonggang, Zhang Shichao, Cheng Debo, *et al*. Spectral clustering based on hypergraph and self-representation [J]. Multimedia Tools and Applications, 2017, 76(16): 17559-17576.

[24] Zhang Shichao, Yang Lifeng, Deng Zhenyun, *et al*. Leverage triple relational structures via low-rank feature reduction for multi-output regression [J]. Multimedia Tools & Applications, 2016, 76(16): 17461-17477.

[25] He Xiaofei, Niyogi P. Locality preserving projections (LPP) [D]. [S.l.]: University of Chicago, 2005.

[26] Zhu Xiaofeng, Zhang Lei, Huang Zi. A sparse embedding and least variance encoding approach to hashing [J]. IEEE Trans on Image Processing, 2014, 23(9): 3737-3750.

[27] Hu Rongyao, Cheng Debo, He Wei, *et al*. Low-rank feature selection for multi-view regression [J]. Multimedia Tools & Applications, 2016, 76(16): 17479-17495.

[28] He Wei, Zhu Xiaofeng, Cheng Debo, *et al*. Low-rank unsupervised graph feature selection via feature self-representation [J]. Multimedia Tools & Applications, 2016, 76(9): 12149-12164.

[29] Spyromitros-Xioufis E, Tsoumakas G, Groves W, *et al*. Multi-target regression via input space expansion: treating targets as inputs [J]. Machine Learning, 2016, 104(1): 55-98.

[30] Karalič A, Bratko I. First order regression [J]. Machine Learning, 1997, 26(2-3): 147-176.

[31] UCI repository of machine learning datasets [EB/OL]. [2016-05-27]. http://archive.ics.uci.edu/ml/.

[32] Cai Xiao, Nie Feipeng, Cai Weidong, *et al*. New graph structured sparsity model for multi-label image annotations [C]//Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2013: 801-808.

[33] Hu Rongyao, Zhu Xiaofeng, Cheng Debo, *et al*. Graph self-representation method for unsupervised feature selection [J]. Neurocomputing, 2016, 220(1): 130-137.

[34] Zhu Pengfei, Zuo Wangmeng, Zhang Lei, *et al*. Unsupervised feature selection by regularized self-representation [J]. Pattern Recognition, 2015, 48(2): 438-446.

[35] Zhu Xiaofeng, Huang Zi, Cui Jiangtao, *et al*. Video-to-shot tag propagation by graph sparse group lasso [J]. IEEE Trans on Multimedia, 2013, 15(3): 633-646.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net