软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006041]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

噪音数据的属性选择算法*

许 航[1], 张师超[1], 吴兆江[1], 李佳烨[2]

[1](中南大学 计算机学院,湖南 长沙 410083)

[2](广西师范大学 计算机科学与信息工程学院,广西 桂林 541004)

通讯作者: 张师超, E-mail: zhangsc@csu.edu.cn

摘 要: 正则化属性选择算法减小噪音数据影响的效果不佳,而且,样本空间的局部结构几乎没有被考虑,在将样本映射到属性子空间后,样本之间的联系与原空间不一致,导致数据挖掘算法的效果不能令人满意.本文提出一个抗噪音属性选择方法,可以有效的解决传统算法的这两个缺陷.该方法首先采用自步学习的训练方式,这不仅能大幅度降低离群点进入训练的可能性,而且有利于模型的快速收敛.然后,采用加入$l_{2,1}$正则项的回归学习器进行嵌入式属性选择,兼顾“求得稀疏解”和“解决过拟合”,使模型更稳健.最后,融合局部保留投影的技术,将其投影矩阵转换成模型的回归参数矩阵,在属性选择的同时保持样本之间的原有局部结构.采用一系列基准数据集合测试该算法,在aCC和aRMSE上的实验结果表明了该属性选择方法的有效性.

关键词: 属性选择;自步学习;局部保留投影

中图法分类号: AP181

中文引用格式: 许航,张师超,吴兆江,李佳烨.噪音数据的属性选择算法.软件学报. http://www.jos.org.cn/1000-9825/6041.htm

英文引用格式: Xu H, Zhang SC, Wu ZJ, Li JY. Feature selection algorithm for noise data. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/6041.htm

Feature Selection Algorithm for Noise Data

XU Hang[1], ZHANG Shi-Chao[1], WU Zhao-Jiang[1], LI Jia-Ye[2]

[1](School of Computer, Central South University, Changsha 410083, China)

[2](School of Computer Science and Information Engineering, Guangxi Normal University, Guilin 541004, China)

Abstract: The regularization feature selection algorithm is not effective in reducing the impact of noisy data. Moreover, the local structure of the sample space is hardly considered. After the samples are mapped to the feature subspace, the relationship between samples is inconsistent with the original space, resulting in unsatisfactory results of the data mining algorithm. This paper proposes an anti-noise feature selection method that can effectively solve these two shortcomings of traditional algorithms. This method first uses a self-paced learning training method, which not only greatly reduces the possibility of outliers entering training, but also facilitates the rapid convergence of the model. Then, a regression learner with regular terms is used to select the embedded features, taking into account the "sparse solution" and "solving over-fitting" to make the model more robust. Finally, the technique of locality preserving projections is integrated, and its projection matrix is transformed into the regression parameter matrix of the model, while maintaining the original local structure between the samples while selecting the features. Some experiments are conducted for evaluating the algorithm with a series of benchmark data sets. Experimental results show the effectiveness of the proposed algorithm in term of the aCC and aRMSE.

Key words: feature selection; self-paced learning; locality preserving projections

* 基金项目: 国家自然科学基金(61836016, 61672177); 中南大学中央高校基本科研业务费专项资金资助(2019zzts964)

Foundation item: National Natural Science Foundation of China (61836016, 61672177); Fundamental Research Funds for the Central Universities of Central South University(2019zzts964)

收稿时间: 2019-12-26; 修改时间: 2020-01-17; 采用时间: 2020-03-27; jos 在线出版时间: 2020-12-02

维度灾难是数据科学中的一个核心问题,属性选择是解决维度灾难的一种重要方法[1-2].在数据挖掘中常见的属性选择主要包括属性加权方法,通过搜索策略[3]寻找最优的属性组合,通过正则化在既定模型中探索那些对模型准确度提升较大的属性[4-5].基于结构化稀疏学习的带正则化项的线性模型是最常用的模型之一,例如,文献[6]中提出对系数矩阵进行 $l_{2,1}$ 正则化约束,在半监督学习过程中进行属性选择.文献[7-8]中引入 l_1 范数,对系数矩阵进行约束选择相关属性.近年来,有些算法在常用模型的基础上进一步考虑其它影响因素,建立性能更为优良的模型.例如,文献[1]将 LASSO 和随机森林结合选择相关的外生变量.文献[9]在高光谱遥感数据挖掘中同时考虑光谱和空间信息,将谱空间特征投影到一个公共的特征空间来学习潜在的子空间模型.文献[10]对系数矩阵同时进行 l_1 范数约束和 $l_{2,1}$ 范数约束,并在此基础上增加低秩约束,得到稀疏鲁棒结果.文献[11]在谱聚类的基础上,定义属性区分度和独立性,以此选择重要属性.文献[12]提出了一种协同正则化稀疏群组 LASSO 算法,允许将辅助信息按照预测因子之间的“组”和“距离”整合到学习任务中.针对数据内部关系在进行子空间学习后可能被破坏的问题,文献[13-14]引入拉普拉斯矩阵,在一定程度上保留原始结构信息.文献[15]采用 k 近邻策略得到样本点的近邻点,并通过相似矩阵和近邻点保留样本空间的邻域结构,然后再进行属性选择.

不过,上述方法通常是加入正则化项来处理噪音数据,其抑制噪音的效果不理想.并且,这些属性选择算法几乎没有考虑数据本身的“难易程度”(“简单”样本为损失函数值小/似然函数值大的样本,也就是置信度高的样本,反之为“复杂样本”),将简单的普适性知识和复杂的专业化知识一概而论,在训练过程中将所有数据(包括噪音点)随机加入训练,使得训练收敛较慢.对此,文献[16]提出将自步学习与属性选择方法结合,考虑数据置信度,达到其去除噪音的目的.其次,在原样本空间中相互之间有联系的样本在属性选择后需要被保持,即,如果两个样本原本是很亲近的关系,则映射到属性子空间时也具有很亲近的关系,这是现有算法很少考虑的问题.针对现有属性选择算法的这两方面不足,本文设计一个新的属性选择算法,同时考虑数据自身的“难易程度”,噪音的多重处理方式,以及样本数据间的局部亲近关系结构.它是一种在训练样本自增长的训练模式下,进行鲁棒稀疏学习并保留样本内部联系的属性选择方法.其主要特点如下:

(1) 采用自步学习的训练模式,实现样本数据的自动增长.在训练过程中,采用“平滑权重”,更加真实的衡量样本的“难易程度”,根据从“简单”到“复杂”的原则逐步选取训练样本,使得模型能够快速收敛.

(2) 在处理噪音时,不仅采用正则化项来减小噪音对模型的影响,而且加入样本的置信度来消除一些噪音样本.

(3) 利用稀疏权值矩阵选择有效属性,并同时采用局部保留投影,使得任意形状的样本空间下都能有效保留样本点的邻域结构.

1 相关理论背景

1.1 符号

给定一个矩阵 $\boldsymbol{X}$,第 i 行表示为 $\boldsymbol{x}_i$,第 j 列表示为 $\boldsymbol{x}^j$,第 i 行 j 列元素表示为 x_{ij}; $\|\boldsymbol{X}\|_2$ 表示矩阵 $\boldsymbol{X}$ 的 l_2 范数,$\|\boldsymbol{X}\|_2=(\sum_{i,j}x_{ij}^2)^{1/2}$; $\|\boldsymbol{X}\|_{2,1}$ 表示表示 $\boldsymbol{X}$ 矩阵的 $l_{2,1}$ 范数,$\|\boldsymbol{X}\|_{2,1}=\sum_i(\sum_j x_{ij}^2)^{1/2}$; $\boldsymbol{X}^{\mathrm{T}}$ 表示矩阵 $\boldsymbol{X}$ 的转置; $tr(\boldsymbol{X})$ 表示矩阵 $\boldsymbol{X}$ 的迹.

1.2 局部保留投影

局部保留投影是一种线性降维算法.在大数据背景下,数据的维度高,但是数据本身内在的有效维数可能远低于数据集呈现的结果.局部保留投影方法(LPP)在降维的同时能保留数据内在的结构,也意味着在低维空间中数据间仍能保留原来的关系[17].

对于每个样本,投影基向量记为 $\boldsymbol{P}$,局部保留投影的目标函数为:

$$\min_{\boldsymbol{P}}\frac{1}{2}\sum_{i,j}^{n}\left(\boldsymbol{x}_i\boldsymbol{P}-\boldsymbol{x}_j\boldsymbol{P}\right)^2\boldsymbol{S} \tag{1}$$

$\boldsymbol{S}$ 为权值系数矩阵,表示样本之间的关系.当样本之间有邻近关系时,对应的权值为非零值; 否则,权值为

0.

一般情况下,利用拉普拉斯矩阵（记作 $\boldsymbol{L}$ ）将目标函数转化为如下形式：

$$\min_{\boldsymbol{P}} tr\left(\boldsymbol{P}^T\boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}\boldsymbol{P}\right) \tag{2}$$

1.3 自步学习

自步学习[18-19]模仿人类学习的方式,从学习简单的知识逐步过渡到学习复杂的知识.自步学习根据置信度或极大似然函数值对样本进行排序,在迭代过程中先选择高置信度或高似然值的“简单”样本,然后逐步加入“复杂”样本.

在自步学习的每一轮迭代中,解决如下的混合整数规划问题：

$$\left(\boldsymbol{W}_{t+1},\boldsymbol{v}_{t+1}\right)=\underset{\boldsymbol{W}\in\boldsymbol{R}^d,\boldsymbol{v}\in\{0,1\}^n}{\operatorname{argmin}}\left(r\left(\boldsymbol{W}\right)+\sum_{i=1}^{n}\boldsymbol{v}_i f\left(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W}\right)-\lambda\sum_{i=1}^{n}\boldsymbol{v}_i\right) \tag{3}$$

其中 r 是关于 $\boldsymbol{W}$ 的正则项, f 是损失函数. $\boldsymbol{v}$ 是自步学习的权重变量,为一个 n 维向量,取值为 0 或 1. v_i 取 1,表示第 i 个样本在本次迭代中被选中,否则未被选中. λ 决定了每次迭代中选中样本的数量, λ 越小,本次迭代就倾向于选择更“简单”的样本,也就是选择更少的样本.

具体来说,上式可变形为：

$$\left(\boldsymbol{W}_{t+1},\boldsymbol{v}_{t+1}\right)=\underset{\boldsymbol{W}\in\boldsymbol{R}^d,\boldsymbol{v}\in\{0,1\}^n}{\operatorname{argmin}}\left(r\left(\boldsymbol{W}\right)+\sum_{i=1}^{n}v_i\left(f\left(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{W}\right)-\lambda\right)\right) \tag{4}$$

由上式可知：对于第 i 个样本,如果损失小于 λ ,为了最小化目标函数, v_i 的值应该为 1；否则, v_i 取值为 0. 在训练过程中, λ 值一直在增大,因此更多的样本被选入训练过程.

1.4 多元回归

回归分析反映事物某一特性随其它因素的变化而变化的规律.多元回归分析研究因变量与两个或两个以上自变量的相关关系问题[20].回归分析因其在统计分析方面的优越性,被广泛的运用于工程技术和社会科学等领域[21].

回归分析问题实际上是通过最小化所有样本的类别预测值与真实值之间的偏差,求出对原始数据拟合程度最好的模型的参数.本文数据集为连续多标签数据集,利用多元回归分析进行属性选择,并验证所提出算法的性能.

2 算法描述和优化

2.1 算法描述

数据集 $\boldsymbol{X}=[\boldsymbol{x}_1;\boldsymbol{x}_2;\ldots;\boldsymbol{x}_n]$, $\boldsymbol{X}\in\boldsymbol{R}^{n\times d}$,n 为样本数,d 为属性数；数据类别 $\boldsymbol{Y}=[\boldsymbol{y}_1;\boldsymbol{y}_2;\ldots;\boldsymbol{y}_n]$, $\boldsymbol{Y}\in\boldsymbol{R}^{n\times c}$,c 为标签数量.

首先给出线性回归学习器的代价函数,本文用均方误差计算损失,将目标函数表示为下列形式：

$$\min_{\boldsymbol{W},\boldsymbol{b}}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{b}\right\|_2^2+\beta\left\|\boldsymbol{W}\right\|_1 \tag{5}$$

代价函数用 l_2 范数计算损失, $\boldsymbol{W}\in\boldsymbol{R}^{d\times c}$ 、 $\boldsymbol{b}\in\boldsymbol{R}^{1\times c}$ 分别为回归参数矩阵、回归参数向量.(5)式对回归参数矩阵进行 l_1 范数约束,经过优化后得到稀疏解,并基于稀疏的回归参数矩阵进行属性选择.然而以这种方式计算的代价函数值容易受异常点、离群点的影响.因为在学习过程中模型会尽可能拟合噪音,从而导致过拟合. l_1 范数一定程度上可以缓解过拟合问题,但一般情况下用 $l_{2,1}$ 范数处理过拟合更加有效.最终,综合考虑求解稀疏解和解决过拟合问题,引入 $l_{2,1}$ 范数,得到目标函数：

$$\min_{\boldsymbol{W},\boldsymbol{b}} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{b} \right\|_2^2 + \beta \left\| \boldsymbol{W} \right\|_{2,1} \tag{6}$$

β 为参数,平衡损失函数和正则化项, β 越大,优化后的 $\boldsymbol{W}$ 矩阵越稀疏.

优化后的 $\boldsymbol{W}$ 矩阵中,稀疏行对应的属性很大程度上是不重要的或者冗余的属性.移除这些属性可以得到新的属性子集,从而实现维度约简.此时 $\boldsymbol{W}$ 近似于新的低维空间的一组基,在这个低维空间中,样本点之间的关系应该与原空间保持一致.为了实现这一目的,借鉴 LPP 的思想,将其中的投影矩阵替换为 $\boldsymbol{W}$ 矩阵.此时,目标函数改进为:

$$\min_{\boldsymbol{W},\boldsymbol{b}} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{b} \right\|_2^2 + \frac{1}{2} \sum_{i,j}^{n} \alpha \left\| \boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{x}_j \boldsymbol{W} \right\|_2^2 s_{ij} + \beta \left\| \boldsymbol{W} \right\|_{2,1} \tag{7}$$

α 为调节参数,调节该项在目标函数中的比重.$\boldsymbol{S}=[\boldsymbol{s}_1;\ \boldsymbol{s}_2;\ldots;\ \boldsymbol{s}_n]$, $\boldsymbol{S} \in \boldsymbol{R}^{n \times n}$, $\boldsymbol{S}$ 为权值矩阵,矩阵中每个元素 $\boldsymbol{s}_{ij}$ 表示样本之间的邻近关系.当 $\boldsymbol{s}_{ij}$ 取非 0 值时,表示第 i 个样本与第 j 个存在邻近关系;否则,表示样本 i 和样本 j 之间不存在邻近关系.$\boldsymbol{s}_{ij}$ 可用 $exp\left(-\frac{\left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_2^2}{2\sigma^2} \right)$ 求得, σ 为参数,不失一般性, σ 值可设为 1.

在大数据背景下,学习器往往要处理庞大的数据集,采用自步学习可以实现样本数据的自动增长以及挖掘模型的快速收敛.在自步学习的迭代过程中,优先选择“简单”的样本,然后逐步加入“复杂”的样本.因此在自步学习过程中,受样本置信度和算法阈值参数的控制,离群点几乎不会被选中,进一步保证了算法的鲁棒性.此时,目标函数改进为:

$$\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{v}} \sum_{i=1}^{n} v_i \left\| \boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{b} \right\|_2^2 + \frac{1}{2} \sum_{i,j}^{n} \alpha \left\| \boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{x}_j \boldsymbol{W} \right\|_2^2 s_{ij} + \beta \left\| \boldsymbol{W} \right\|_{2,1} - \lambda \left\| \boldsymbol{v} \right\|_1 \tag{8}$$

$\boldsymbol{v} \in \boldsymbol{R}^{1 \times n}$, $\boldsymbol{v}$ 为自步学习权重变量,$\boldsymbol{v}=[v_1,v_2,\ldots,v_n]$. λ 为阈值参数, $\lambda > 0$ 为约束条件. v_i 的取值由 λ 值和损失函数值共同决定:当 $v_i = 1$ 时,表示第 i 个样本被选入训练过程;否则,表示在下次迭代过程,暂时不考虑第 i 个样本.但是噪音不是均匀分布在数据集中,绝对化的断定某一个样本是“简单”的或是“复杂”的是不合理的.因此,根据文献[22]改变关于 $\boldsymbol{v}$ 的正则项,将自步学习中的“硬权重”改为“平滑权重”,目标函数如下:

$$\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{v}} \sum_{i=1}^{n} v_i \left\| \boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{b} \right\|_2^2 + \frac{1}{2} \sum_{i,j}^{n} \alpha \left\| \boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{x}_j \boldsymbol{W} \right\|_2^2 s_{ij} + \beta \left\| \boldsymbol{W} \right\|_{2,1} + \lambda \left(\frac{1}{2} \left\| \boldsymbol{v} \right\|_2^2 - \left\| \boldsymbol{v} \right\|_1 \right) \tag{9}$$

算法 1:多目的属性选择

输入:训练样本 $\boldsymbol{X} \in \boldsymbol{R}^{n \times d}$, $\boldsymbol{Y} \in \boldsymbol{R}^{n \times c}$,参数 α, β, λ

输出:平均相关系数 aCC,平均均方根误差 aRMSE:

1 采用算法 2 求解稀疏回归参数矩阵 $\boldsymbol{W} \in \boldsymbol{R}^{d \times c}$

2 在原始数据集上,根据得到的 $\boldsymbol{W}$ 进行属性选择,得到新的数据集

3 在新的数据集上用支持向量回归(SVR)、核岭回归(KRR)进行回归分析,验证属性选择算法的性能

2.2 算法优化

本节对目标函数进行优化,采用机器学习中常用的求解方法:交替固定变量求导,对目标函数(9)进行求解.

(1) 为了方便计算,首先将目标函数进行化简.

1) 对目标函数第一项有如下变换:

首先设置如下变量:

$$\boldsymbol{Q} = \left[\sqrt{v_1}\boldsymbol{y}_1; \sqrt{v_2}\boldsymbol{y}_2; \ldots; \sqrt{v_n}\boldsymbol{y}_n \right] \in \boldsymbol{R}^{n \times c}$$

$$\boldsymbol{G} = \left[\sqrt{v_1}\boldsymbol{x}_1; \sqrt{v_2}\boldsymbol{x}_2; \ldots; \sqrt{v_n}\boldsymbol{x}_n \right] \in \boldsymbol{R}^{n \times d}$$

$$U=\begin{pmatrix}\sqrt{v_1}\\ \sqrt{v_2}\\ \vdots\\ \sqrt{v_n}\end{pmatrix}\in \boldsymbol{R}^{n\times 1}$$

则第一项通过如下过程转化：

$$\sum_{i=1}^{n}v_i\left\|\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{b}\right\|_2^2=\sum_{i=1}^{n}\left\|\sqrt{v_i}\left(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{b}\right)\right\|_2^2=\left\|\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right\|_2^2 \tag{10}$$

2) 对目标函数第二项有如下变换：

首先设 $\boldsymbol{D}=diag\left(\left[\sum_{j=1}^{n}s_{1j},\sum_{j=1}^{n}s_{2j},\ldots,\sum_{j=1}^{n}s_{nj}\right]\right)\in\boldsymbol{R}^{n\times n}$,进行下列转化：

$$\begin{aligned}&\frac{1}{2}\sum_{i,j}^{n}\alpha\left\|\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{x}_j\boldsymbol{W}\right\|_2^2s_{ij}=\\&\alpha\sum_{i,j}^{n}\left(\boldsymbol{W}^T\boldsymbol{x}_i{}^Ts_{ij}\boldsymbol{x}_i\boldsymbol{W}-\boldsymbol{W}^T\boldsymbol{x}_i{}^Ts_{ij}\boldsymbol{x}_j\boldsymbol{W}\right)=\\&\alpha\left(\sum_{i}^{n}\boldsymbol{W}^T\boldsymbol{x}_i{}^Td_{ii}\boldsymbol{x}_j\boldsymbol{W}-\sum_{i,j}^{n}\boldsymbol{W}^T\boldsymbol{x}_i{}^Ts_{ij}\boldsymbol{x}_j\boldsymbol{W}\right)=\\&\alpha tr\left(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}\boldsymbol{W}-\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X}\boldsymbol{W}\right)=\\&\alpha tr\left(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right)\end{aligned} \tag{11}$$

$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$, $\boldsymbol{L}$ 为拉普拉斯矩阵.

最终,目标函数转化为：

$$\min_{\boldsymbol{W},\boldsymbol{B},\boldsymbol{v}}\left\|\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right\|_2^2+\alpha tr\left(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right)+\beta\left\|\boldsymbol{W}\right\|_{2,1}+\lambda(\frac{1}{2}\left\|\boldsymbol{v}\right\|_2^2-\left\|\boldsymbol{v}\right\|_1) \tag{12}$$

(2) 固定 $\boldsymbol{v}$、$\boldsymbol{W}$ 后优化 $\boldsymbol{b}$,此时目标函数为:

$$\min_{\boldsymbol{b}}\left\|\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right\|_2^2 \tag{13}$$

对 $\boldsymbol{b}$ 求导,令导数为 0,结果为：

$$\boldsymbol{b}=(\boldsymbol{U}^T\boldsymbol{U})^{-1}(\boldsymbol{U}^T\boldsymbol{Q}-\boldsymbol{U}^T\boldsymbol{G}\boldsymbol{W}) \tag{14}$$

(3) 固定 $\boldsymbol{v}$,$\boldsymbol{b}$ 后优化 $\boldsymbol{W}$,此时目标函数为：

$$\min_{\boldsymbol{W}}\left\|\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right\|_2^2+\alpha tr\left(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right)+\beta\left\|\boldsymbol{W}\right\|_{2,1} \tag{15}$$

定义一个对角矩阵 $\boldsymbol{O}$ ：

$$\boldsymbol{O}_{ii}=\frac{1}{2\left\|\boldsymbol{w}_i\right\|_2}$$

得 $2tr\left(\boldsymbol{W}^T\boldsymbol{O}\boldsymbol{W}\right)=\left\|\boldsymbol{W}\right\|_{2,1}$,“2”与“ β ”都是系数,将两者合并在一起,目标函数转换为：

$$\min_{\boldsymbol{W}}\left(\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right)^T\left(\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right)+\alpha tr\left(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}\boldsymbol{W}\right)+\beta tr\left(\boldsymbol{W}^T\boldsymbol{O}\boldsymbol{W}\right) \tag{16}$$

对（12）中 $\boldsymbol{W}$ 求导,令导数为 0,结果为：

$$\boldsymbol{W}=\left(\boldsymbol{G}^{T}\boldsymbol{G}+\alpha\boldsymbol{X}^{T}\boldsymbol{L}\boldsymbol{X}+\beta\boldsymbol{O}\right)^{-1}\left(\boldsymbol{G}^{T}\boldsymbol{Q}-\boldsymbol{G}^{T}\boldsymbol{U}\boldsymbol{b}\right) \tag{17}$$

(4) 固定 $\boldsymbol{B},\boldsymbol{W}$ 后优化 $\boldsymbol{v}$,此时目标函数为：

$$\min_{\boldsymbol{v}}\left\|\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right\|_2^2+\lambda(\frac{1}{2}\left\|\boldsymbol{v}\right\|_2^2-\left\|\boldsymbol{v}\right\|_1) \tag{18}$$

$\boldsymbol{v}$ 的取值根据损失函数和参数 λ 决定：

$$v_i=\begin{cases}-\dfrac{l}{\lambda}+1; l<\lambda\\ 0; l\geq\lambda\end{cases} \tag{19}$$

l 表示损失函数 $\boldsymbol{F}\left(\boldsymbol{W},\boldsymbol{x}_i,\boldsymbol{y}_i\right)$ 的值： $\left\|\boldsymbol{Q}-\boldsymbol{G}\boldsymbol{W}-\boldsymbol{U}\boldsymbol{b}\right\|_2^2$

算法 2：求解稀疏回归参数矩阵

输入：训练样本 $\boldsymbol{X}\in\boldsymbol{R}^{n\times d}$, $\boldsymbol{Y}\in\boldsymbol{R}^{n\times c}$,参数 α，β，λ

输出：稀疏回归参数矩阵 $\boldsymbol{W}\in\boldsymbol{R}^{d\times c}$

1 t=1//t 为迭代次数

2 初始化参数 α，β，λ, $\boldsymbol{W}$,$\boldsymbol{b},\boldsymbol{O}$

3 通过（14）式计算 $\boldsymbol{b}$

4 通过(17)式计算 $\boldsymbol{W}$

5 根据 $\boldsymbol{W}$ 更新 $\boldsymbol{O}$

6 改变阈值 λ,根据 λ,更新 ν

7 t=t+1

8 重复 3~7,直到目标函数收敛或 t 达到迭代次数时结束

3 实验

3.1 数据集和对比方法

因为本文提出的算法为属性选择算法,实验将选择属性数目较多的数据集进行研究.实验使用了来自 Mulan 的 6 个数据集:Atp1d[23],Atp7d[23],Scm1d[23],Oes10[23],Oes97[23],Rf2[23],其信息如下：

Atp1d、Atp7d：关于价格预测的机票价格数据集.数据按时间顺序排列.该数据集中的每个样本表示一组来自特定观测日期和出发日期的观测结果.每个样本的输入变量可能是对预测特定起飞日期的机票价格有用的值.每个样本的目标变量是 6 个目标飞行偏好的次日(Atp1d)价格或未来 7 天观察到的最低价格(Atp7d).

Oes10、Oes97：美国劳工统计局从 1997 年(Oes97 年)和 2010 年(Oes10 年) 汇编的年度职业就业调查数据集.每一行提供了一个特定的大都市地区中众多就业类型的全职员工的估计数量.

Scm1d：2010 年供应链管理(TAC SCM)锦标赛中贸易代理比赛的供应链管理数据集.该数据集包含 16 个回归目标,每个目标对应每个模拟产品第二天的平均价格.

Rf2：美国国家气象局在特定地点采集的对未来 48 小时河网流量进行预测的河流流量数据集.该数据集包含了美国密西西比河网络的 8 个站点的每小时流量观测数据.

所有数据集的样本数、属性数、标签数详情如下：

Table 1 Experimental data set information

表 1 实验数据集信息

数据集	样本数	属性数	标签数
Atp1d	337	411	6
Atp7d	296	411	6
Scm1d	9803	280	16

Oes10	403	298	16
Oes97	334	263	16
Rf2	9125	576	8

实验采用了 4 种效果比较好的属性选择算法进行对比,并将属性选择前的回归结果作为参照.CSFS[6]是一种凸半监督属性选择算法,与某些算法相比它不需要构造图,不需要进行复杂的特征分解.LSG21[24]算法通过建立一个图结构稀疏模型来进行稀疏属性选择.SLRR[25]通过构造系数矩阵的低秩稀疏结构进行属性选择.URAFS[26]算法是将广义非相关约束和流形学习结合的无监督特征选择方法,能有效的排除冗余属性.

3.2 实验设置

实验首先利用属性选择方法选取属性,由于本算法为典型的嵌入式属性选择方法,学习器为多元线性回归模型,所以在选取的属性子集上进行回归任务来分析算法性能时,为了避免巧合性,采用其它回归分析的算法（本文采用的 SVR 和 KRR）.

实验采用 10 折交叉验证进行训练/测试,对于参数 α，β ,取值范围设置在$[0.3*10^{-3},3*10^{3}]$内.

对于数据集 Rf2,其存在少量的有缺失值的数据,因为本文算法不考虑缺失值的影响,所以在实验中剔除缺失数据.其次,它的数据量较大,采用 Matlab 并行计算工具箱进行实验,速度依然非常慢.为了解决此问题,在实际操作中对交叉验证划分的互斥子集进行 2~4 次划分,然后选取最终的样本进行实验.（交叉验证法采用的分层采样,能尽量保持数据分布的一致性）.对于数据量同样大的 Scm1d 数据集,采取类似的处理方式.

实验采用平均相关系数(average correlation coefficient, aCC) 和平均均方根误差(average Root Mean Square Error, aRMSE)度量模型的性能.实验结果用指标±标准差表示,且都采用百分数形式.其中,平均相关系数反映了响应变量和自变量之间的线性相关性所引起的波动大小,值越大,线性回归越好,表示响应变量与自变量（本文的自变量从属性选择后的新数据集获取）之间的线性相关程度越高.平均均方根误差用来衡量预测值与真实值之间的误差,值越小越好,表示模型准确度越高.简而言之,这两个指标从两个方面衡量算法的性能：

(1) 属性选择后的属性与响应变量的相关性（属性选择算法应该尽量选择与研究问题相关的特征）.

(2) 属性选择后的属性子集对回归模型预测正确性的影响.

3.3 实验结果和分析

实验采用 10 折交叉验证进行训练/测试,比较进行属性选择前后的效果,表 2（使用 SVR 验证属性选择算法）、表 3（使用 KRR 验证属性选择算法）为使用属性选择前以及使用各类算法进行属性选择后在 aCC 指标上的比较结果.

Table 2 aCC comparison results（SVR）

表 2 aCC 对比结果（SVR）

数据集 / 算法	Atp1d	Atp7d	Scm1d	Oes10	Oes97	Rf2
LSG21	92.75±2.27	85.61±8.06	90.35±3.82	93.00±4.53	86.68±4.82	89.36±2.03
SLRR	92.86±3.06	84.05±8.47	89.95±4.17	92.83±4.48	84.20±9.49	89.41±2.50
URAFS	92.42±3.90	85.28±8.53	90.12±2.31	91.75±7.26	85.48±6.88	89.13±2.82
Proposed	94.40±1.48	89.04±3.44	91.42±2.96	94.29±2.61	89.55±3.76	91.58±2.81

No Feature Selection	91.73±3.77	81.97±8.60	89.39±2.87	90.50±5.00	83.28±10.22	88.55±1.88

从表 2 中可以看出,在这 6 个数据集上,进行属性选择后的实验结果比直接进行回归时更好.进行属性选择时,本算法的 aCC 高于对比算法.具体来说,在数据集 Atp1d 上,本算法的 aCC 平均值为 94.40%,分别比其它对比算法高出 1.13%,1.65%,1.54%,1.98%.在数据集 Atp7d 上,本算法取得了最高的 aCC 平均值 89.04%,分别比其它对比算法高出 3.72%,3.43%,4.99%,3.76%.在数据集 Scm1d 上,本算法的结果最好,分别比其它对比算法高出 1.25%,1.07%,1.47%,1.30%.在数据集 Oes10 上,本算法的 aCC 平均值为 94.29%,分别比其它对比算法高出 1.31%,1.29%,1.46%,2.54%.在数据集 Oes97 上,本算法的 aCC 平均值分别比其它对比算法高出 2.50%,2.87%,5.35%,4.07%.在数据集 Rf2 上,本算法的 aCC 平均值为 91.58%,分别比其它对比算法高出 2.54%,2.22%,2.17%,2.45%.

Table 3 aCC comparison results（KRR）

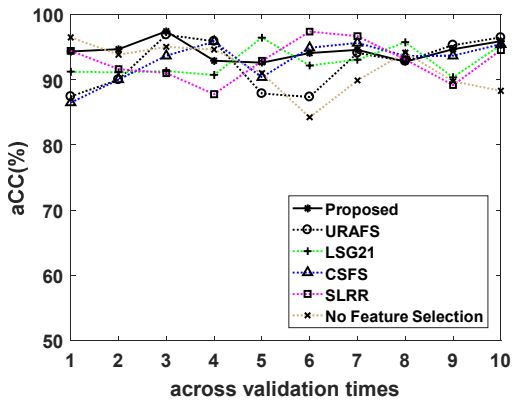
表 3 aCC 对比结果（KRR）

数据集 / 算法	Atp1d	Atp7d	Scm1d	Oes10	Oes97	Rf2
CSFS	90.90±3.82	81.73±14.44	85.32±4.95	64.62±25.52	51.31±26.27	64.06±9.58
LSG21	90.79±3.78	80.26±16.20	86.67±4.14	67.99±21.62	54.52±21.04	63.28±10.02
SLRR	91.43±4.17	84.88±8.71	85.59±6.26	64.53±23.08	55.60±25.54	63.32±11.52
URAFS	91.18±3.30	82.17±11.03	84.41±4.93	61.96±24.96	53.00±21.71	64.50±14.91
Proposed	92.75±1.59	85.94±7.24	88.05±3.27	70.01±22.94	59.91±21.30	67.04±12.99
No Feature Selection	90.49±3.37	79.73±16.13	84.29±6.82	58.93±27.31	49.95±19.98	57.40±7.39

从表 3 可以看出,由于两个回归算法性能的差别,在这 6 个数据集上,用 SVR 进行回归的效果要更好.但是,在同一种回归算法、同一数据集的实验环境中,属性选择后的结果更好,且本算法的指标值略高于其它对比算法.

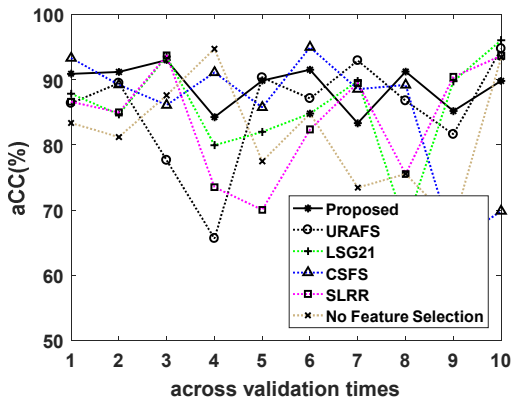
以上实验结果显示在本算法获取的子集上训练的回归模型是最稳定的,意味着本算法选取的属性与响应变量的相关性更强,更贴合相应的数据集所代表的研究内容.这是因为本算法保留了样本的原始内部结构,使得样本在子空间上的结构分布跟原始空间一致.而其他算法忽视了这一点,此时映射到子空间的样本之间的联系（比如邻近关系）可能截然相反,从而使得这种属性子集代表的研究内容与原本研究的内容产生了出入.

图 1 表示采用 SVR 进行回归时,在 10 折交叉验证过程中的每一次训练结果.



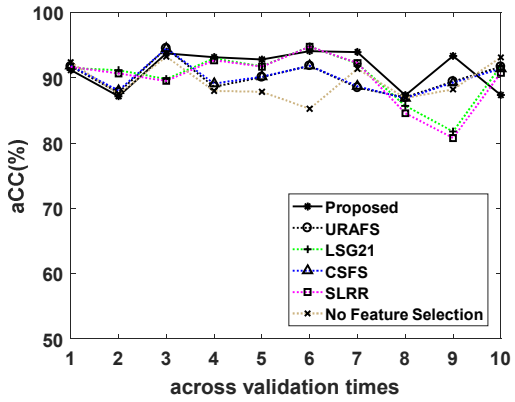
(a) Data set Atp1d

(a) 数据集 Atp1d



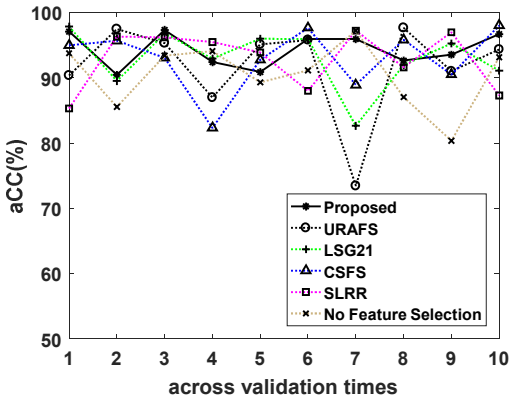
(b) Data set Atp7d

(b) 数据集 Atp7d



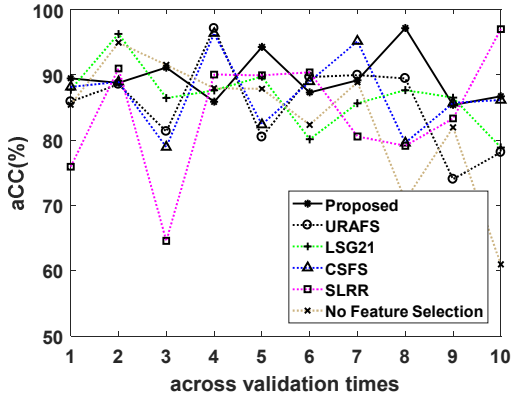
(c) Data set Scm1d

(c) 数据集 Scm1d



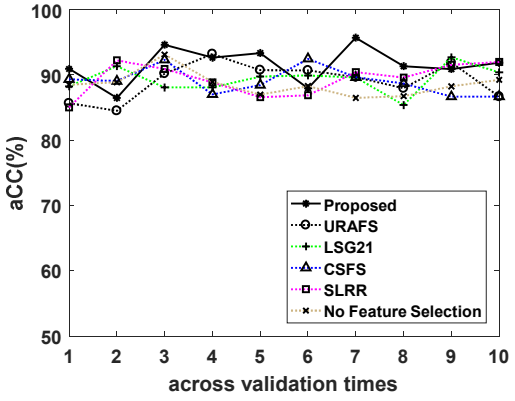
(d) Data set Oes10

(d) 数据集 Oes10



(e) Data set Oes97

(e) 数据集 Oes97



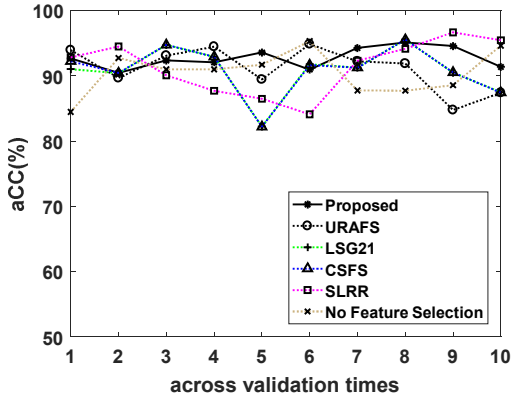
(f) Data set Rf2

(f) 数据集 Rf2

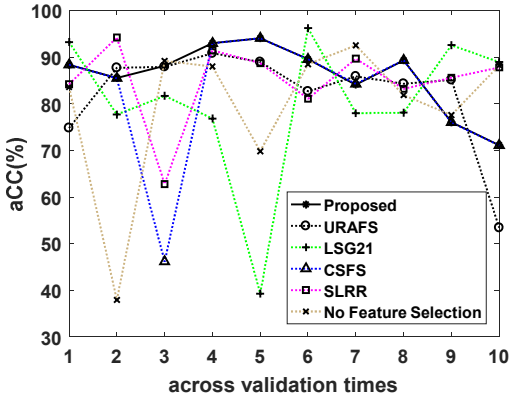
Fig.1 Experimental result（SVR）

图 1 实验结果图（SVR）

图 2 表示采用 KRR 进行回归时,在 10 折交叉验证过程中的每一次训练结果.

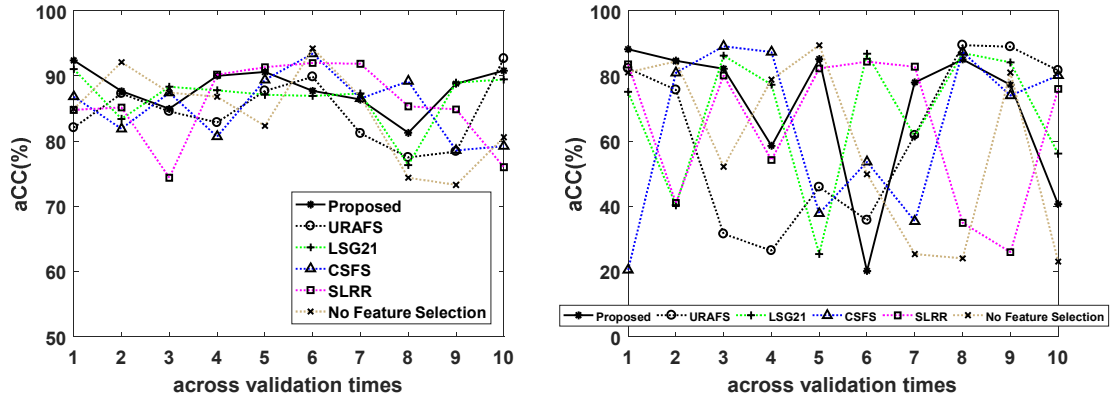


(a) Data set Atp1d



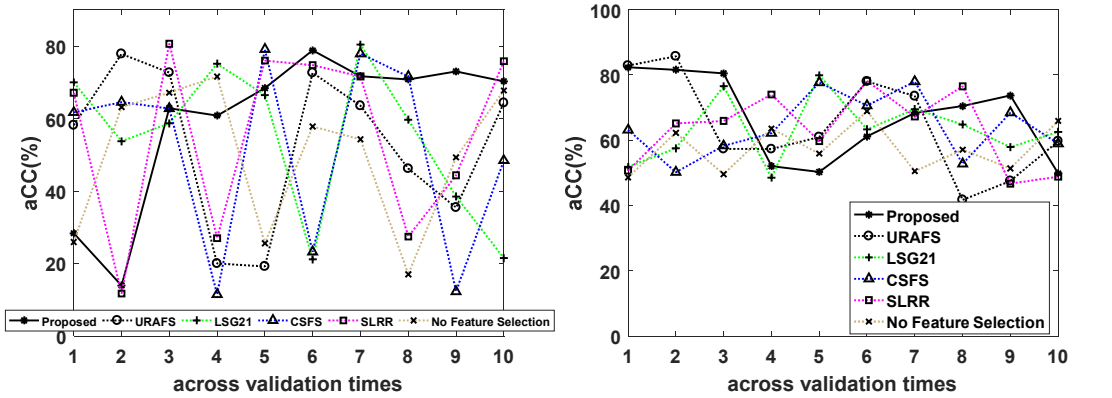
(b) Data set Atp7d

(a) 数据集 Atp1d (b) 数据集 Atp7d



(c) Data set Scm1d (d) Data set Oes10

(c) 数据集 Scm1d (d) 数据集 Oes10



(e) Data set Oes97 (f) Data set Rf2

(e) 数据集 Oes97 (f) 数据集 Rf2

Fig.2 Experimental result（KRR）

图 2 实验结果图（KRR）

从两图中可以看出,用不同的算法进行回归任务时,进行属性选择确实能提升学习任务的效果,因为属性选择过程中尽可能的剔除了无关属性和冗余属性.除此之外,还可以看出,在进行属性选择时,由于数据划分的随机性和调节参数的不同设置,本算法在每一个数据集的每一折训练结果中,并不都是最好的,但是在大多数情况下,本算法的结果优于对比算法,所以最终的平均相关系数和平均均方根误差呈现最优值.而且从图像的波动程度来看,其他算法对应的折线震荡幅度都很大,说明它们的模型相对而言不是很稳定,进一步证明了本算法对应的回归模型鲁棒性高.

本算法与其它算法在 aRMSE 指标上的比较结果如表 4(SVR 验证)、表 5(KRR 验证)所示.表中的 Average 为在同一个数据集上,所有算法 aRMSE 的平均值.

Table 4 aRMSE comparison results（SVR）

表 4 aRMSE 对比结果（SVR）

数据集 / 算法	Atp1d	Atp7d	Scm1d	Oes10	Oes97	Rf2

CSFS	0.70±0.11	0.73±0.06	0.31±0.04	1.30±0.39	1.62±0.46	0.63±0.07
LSG21	0.70±0.12	0.72±0.16	0.30±0.07	1.29±0.42	1.64±0.49	0.62±0.06
SLRR	0.70±0.14	0.72±0.10	0.31±0.07	1.27±0.50	1.64±0.52	0.64±0.06
URAFS	0.69±0.12	0.73±0.15	0.31±0.04	1.29±0.41	1.66±0.45	0.63±0.08
Proposed	0.65±0.11	0.70±0.10	0.29±0.04	1.25±0.39	1.61±0.43	0.56±0.07
No Feature Selection	0.74±0.15	0.83±0.12	0.32±0.04	1.85±1.46	1.99±1.01	0.68±0.06
Average	0.70±0.13	0.74±0.12	0.31±0.05	1.38±0.60	1.69±0.56	0.63±0.07

Table 5 aRMSE comparison results（KRR）

表 5 aRMSE 对比结果（KRR）

数据集 / 算法	Atp1d	Atp7d	Scm1d	Oes10	Oes97	Rf2
CSFS	0.78±0.10	0.80±0.25	0.38±0.06	3.02±1.95	3.65±2.15	1.44±0.22
LSG21	0.79±0.11	0.83±0.28	0.37±0.09	2.89±2.49	3.46±1.85	1.41±0.23
SLRR	0.78±0.17	0.76±0.10	0.38±0.09	3.03±2.11	3.32±2.13	1.43±0.31
URAFS	0.80±0.17	0.84±0.23	0.40±0.06	3.10±1.81	3.61±2.00	1.40±0.36
Proposed	0.73±0.09	0.74±0.13	0.35±0.06	2.73±2.16	3.11±2.30	1.32±0.32
No Feature Selection	0.80±0.09	0.86±0.26	0.39±0.08	3.23±1.86	3.67±2.18	1.59±0.19
Average	0.78±0.12	0.81±0.21	0.38±0.07	3.00±2.06	3.47±2.10	1.43±0.27

从两表中可以看出,无论在哪种回归任务中,进行属性选择后的模型正确性要更高.进行属性选择后,本算法在各个数据集上的 aRMSE 值均为最小,比如在表 4 中,其值分别为 0.65%±0.11%,0.70%±0.10%,0.29%±0.04%,1.25%±0.39%,1.61%±0.43%,0.56%±0.07%,低于各个数据集上所有算法 aRMSE 值的平均值,在表 5 中情况一致.这反映出本算法对应的回归模型正确性高,实际上表明学习器在本算法选择的属性子集上更容易完成学习任务.这是因为本算法通过多种方式解决了噪音问题,即不仅通过正则项修正了噪音带来的影响,还通过考虑数据自身相对于模型的置信度,采用了特殊的训练方式来避免大部分噪音点进入训练.

4 结束语

本文提出了一种在样本自增长的训练模式下,考虑数据间固有联系并进行稀疏学习的属性选择算法.本算法通过考虑样本的置信度判断样本是否加入下一次迭代过程,在整个优化过程中优先选择置信度高的样本进行训练,同时使训练样本数量自动增长.在训练过程中,在选择非冗余属性的同时保留数据间的相关关系.并且通过特殊的训练模式和 $l_{2,1}$ 正则化项避免离群点等噪音对模型的影响,使整个算法更具有鲁棒性.与现有算法比较的实验结果表明,本算法能有效地选择重要属性.在今后的研究中,可考虑通过改变正则化项或采用稀疏的低秩约束扩展模型,以达到更好的效果.

References:

[1] Ludwig N, Feuerriegel S, Neumann D. Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. Journal of Decision Systems, 2015, 24(1): 19-36.

[2] 刘艺,曹建军,刁兴春,周星.特征选择稳定性研究综述.软件学报,2018,29(09):2559-2579.

[3] 初蓓,李占山,张梦林,于海鸿.基于森林优化特征选择算法的改进研究.软件学报,2018,29(09):2547-2558.

[4] 刘飞飞.特征选择算法及应用综述.办公自动化,2018,23(21):47-49.

[5] Zhu X, Li X, Zhang S. Block-row sparse multiview multilabel learning for image classification. IEEE transactions on cybernetics, 2015, 46(2): 450-461.

[6] Chang X, Nie F, Yang Y, et al. A convex formulation for semi-supervised multi-label feature selection. Twenty-eighth AAAI conference on artificial intelligence. 2014.

[7] Fan J, Li R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. arXiv preprint math/0602133, 2006.

[8] Xie Z, Xu Y. Sparse group LASSO based uncertain feature selection. International Journal of Machine Learning and Cybernetics, 2014, 5(2): 201-210.

[9] Zhang L, Zhang Q, Du B, et al. Simultaneous spectral-spatial feature selection and extraction for hyperspectral images. IEEE Transactions on Cybernetics, 2016, 48(1): 16-28.

[10] Wang H, Nie F, Huang H, et al. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. 2011 International Conference on Computer Vision. IEEE, 2011: 557-562.

[11] 谢娟英,丁丽娟,王明钊.基于谱聚类的无监督特征选择算法.软件学报[2020-02-02].https://doi.org/10.13328/j.cnki.jos.005927.

[12] Santos P L A, Imangaliyev S, Schutte K, et al. Feature selection via co-regularized sparse-group Lasso. International Workshop on Machine Learning, Optimization, and Big Data. Springer, Cham, 2016: 118-131.

[13] Zhu X, Li X, Zhang S, et al. Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE transactions on neural networks and learning systems, 2016, 28(6): 1263-1275.

[14] 孙圣姿, 万源, 曾成. 自适应嵌入的半监督多视角特征降维. 计算机应用, 2018: 0-0

[15] 刘艳芳,叶东毅.基于邻域保持学习的无监督特征选择算法.模式识别与人工智能,2018,31(12):1096-1102.

[16] 甘江璋. 基于自步学习和鲁棒估计的属性选择算法研究[博士论文].广西师范大学,2019.

[17] He X, Niyogi P. Locality preserving projections. Advances in neural information processing systems. 2004: 153-160.

[18] Zhao Q, Meng D, Jiang L, et al. Self-paced learning for matrix factorization. Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[19] Lin L, Wang K, Meng D, et al. Active self-paced learning for cost-effective and progressive face identification. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(1): 7-19.

[20] DeForest D K, Brix K V, Tear L M, et al. Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. Environmental toxicology and chemistry, 2018, 37(1): 80-90.

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

[22] Jiang L, Meng D, Zhao Q, et al. Self-paced curriculum learning. Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

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

[24] Cai X, Nie F, Cai W, et al. New graph structured sparsity model for multi-label image annotations. Proceedings of the IEEE International Conference on Computer Vision. 2013: 801-808.

[25] Cai X, Ding C, Nie F, et al. On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 1124-1132.

[26] Li X, Zhang H, Zhang R, et al. Generalized uncorrelated regression with adaptive graph for unsupervised feature selection. IEEE transactions on neural networks and learning systems, 2018, 30(5): 1587-1595.

附中文参考文献:

[2] 刘艺,曹建军,刁兴春,周星.特征选择稳定性研究综述.软件学报,2018,29(09):2559-2579.

[3] 初蓓,李占山,张梦林,于海鸿.基于森林优化特征选择算法的改进研究.软件学报,2018,29(09):2547-2558.

[4] 刘飞飞.特征选择算法及应用综述.办公自动化,2018,23(21):47-49.

[11] 谢娟英,丁丽娟,王明钊.基于谱聚类的无监督特征选择算法.软件学报[2020-02-02].https://doi.org/10.13328/j.cnki.jos.005927.

[14] 孙圣姿, 万源, 曾成. 自适应嵌入的半监督多视角特征降维. 计算机应用, 2018: 0-0

[15] 刘艳芳,叶东毅.基于邻域保持学习的无监督特征选择算法.模式识别与人工智能,2018,31(12):1096-1102.

[16] 甘江璋. 基于自步学习和鲁棒估计的属性选择算法研究[博士论文].广西师范大学,2019.