基于标签特定特征的多目标回归稀疏集成方法

刘洪涛[1,2],李 航[1],王 进[1],李鸽鸽[1,2]
(1. 重庆邮电大学计算机学院,重庆 400065;2. 重庆邮电大学网络智能与网络技术研究中心,重庆 400065)

摘 要: 多目标回归学习是指同时学习多个相关的回归任务,其主要挑战来自于对输入要素和输出目标变量之间的基础关系进行建模以及对目标间的相关性进行探索.针对这两个挑战,本文提出了一种基于标签特定特征的多目标回归稀疏集成方法,通过探索目标间的相关性,为每个目标构建其独特的标签特定特征,提高算法整体的预测精度;同时设计一种稀疏性聚合函数对不同的回归方法进行集成,从而处理输入与输出间的复杂关系.在 18 个数据集上与有代表性的多目标回归方法进行对比实验,充分证明了本文方法的有效性与竞争性.

关键词: 多目标回归; 稀疏集成; 标签特定特征; 目标间关联
中图分类号: TP391, TP399 **文献标识码:** A **文章编号:** 0372-2112 (2020)05-0906-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.05.010

Multi-Target Regression via Sparse Integration and Label-Specific Features

LIU Hong-tao[1,2], LI Hang[1], WANG Jin[1], LI Ge-ge[1,2]
(1. *School of Computer, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;
2. *Chongqing University of Posts and Telecommunications Network Intelligence and Network Technology Research Center, Chongqing* 400065, *China*)

Abstract: Multi-target regression (MTR) refers to learning multiple relevant regression tasks simultaneously. The main challenges of multi-target regression arise from modeling the underlying relationships between input features and output target variables as well as exploring inter-target correlations. In this paper, we propose a multi-target regression method via sparse integration and label-specific features (SI-LSF) that utilizes inter-target correlations to improve the overall prediction accuracy by constructing label-specific features and deals with the input-output relationships through sparse integration of various regression models. Extensive experimental evaluation on 18 benchmark datasets demonstrates that our proposed method can achieve competitive performance against representative state-of-the-art multi-target regression methods, which shows the great effectiveness in dealing with multivariate prediction.

Key words: multi-target regression; sparse integration; label specific features; inter-target correlations

1 引言

多目标回归(Multi-Target Regression, MTR)是使用一组公共的输入变量预测多个连续输出变量的回归任务[1].由于多目标回归具有同时预测多个相关目标的能力,现已广泛应用于各个实际领域中[2].此外,还可应用于计算机视觉和图像分析等领域[3].

多目标回归的主要挑战是[4,5]:(1)对输入数据与输出值之间的复杂关系进行建模;(2)探索目标间相关性,通过学习一个输入变量间线性组合的函数预测输出变量的值[6].它通过将原始输入空间映射到具有非线性内核的高维甚至无限维的内核空间来解开复杂的非线性输入与输出关系[7].因此在实际应用中利用内核函数处理输入特征与输出目标间的复杂关系的方式可能不够灵活.

针对通过探索多个目标之间依赖关系来提高预测准确性[4],为扩展特征空间上的每个目标建立单独的回归模型,并将其它目标的预测值视为附加特征[4].但训练集和预测集之间的附加特征值的差异可能导致预测性能的急剧退化.

针对目前多目标回归面临的挑战,本文提出基于标签特定特征的多目标回归稀疏集成方法(Multi-Target Regression via Sparse Integration and Label-Specific features, SI-LSF),有效地提高算法的预测性能以及提高处理多目标数据的灵活性.首先,对每个目标使用单一的回归方法进行学习与预测将其预测值作为附加特征扩展到原始输入特征空间中.进而,对扩展后的特征空间

收稿日期:2019-03-27;修回日期:2019-11-08;责任编辑:马兰英
基金项目:国家自然科学基金(No.41571401);国家社会科学基金(No.18BGL266)

进行学习,为每个目标学习其标签特定特征. 最后设计一种稀疏性聚合函数,对复杂的输入与输出关系建模,使其既能处理线性关系,也能处理非线性关系. 通过结合稀疏集成和标签特定特征,可以在单一框架中同时解决多目标回归的主要挑战.

本文提出的 SI-LSF 方法具有以下优点.

(1) 改进了预测性能. 本文提出了标签特定特征,考虑了来自原始特征空间的输入特征和目标之间以及目标与目标间的相关性,极大地改善了预测性能.

(2) 提升了处理输入与输出关系的灵活性. 本文提出了基于标签特定特征的稀疏集成方法,自动选择适当的回归模型将输入要素映射到相应的输出目标中.

2　相关工作

多目标回归的研究在以前主要集中在特征方面,例如简单地学习多个任务的特征[8,9]或单独探索特定的任务结构[10,11]. 大多数的方法利用在线性回归模型探索目标间相关性,但缺乏了处理高维输入和多个目标之间的非线性关系的能力. 文献[11]提出了一种多输出回归模型(Multiple-Output Regression,MROTS),该方法利用了潜在模型参数的协方差结构和输出空间的条件协方差结构,并推广了多变量回归模型与协方差估计[12]和线性关系多任务学习[13]. 文献[14]提出了一种被称为校准多元回归(Calibrated Multivariate Regression,CMR)的线性回归模型,可以解决不同任务的不同水平的噪声.

为了处理输入要素与输出目标的非线性关系,输出内核学习(Output Kernel Learning,OKL)算法[15]学习了多个目标的半确定相似性矩阵,即输出内核,然而它不能完全挖掘出目标间的相关性,例如负相关[13]. 文献[16]通过假设所有任务可以聚类成不相交的组,开发了基于聚类的多目标学习(Clustered Multi-Target Learning,CMTL),从训练数据中学习底层聚类结构,探索目标间的相关性,但该算法需要指定类簇的数量,不适用于实际场景.

3　基于标签特定特征的多目标回归稀疏集成算法

3.1　符号定义

设 $\boldsymbol{X}\subset R^m$ 是由 m 维特征向量 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m$ 构成的输入特征空间,$\boldsymbol{Y}\subset R^d$ 表示有 d 个目标变量 $Y_1,Y_2,\cdots,Y_d$ 的输出空间. 用$(\boldsymbol{x}^{(l)},\boldsymbol{y}^{(l)})$表示一个样本,其中 $l\in\{1,\cdots,n\}$,输入向量用 $\boldsymbol{x}^{(l)}=(x_1^{(l)},\cdots,x_m^{(l)})$表示,$\boldsymbol{y}^{(l)}=(y_1^{(l)},\cdots,y_d^{(l)})$表示为输出向量且 $l\in\{1,\cdots,n\}$,训练集 $D=\{(\boldsymbol{x}^{(1)},\boldsymbol{y}^{(1)}),\cdots,(\boldsymbol{x}^{(n)},\boldsymbol{y}^{(n)})\}$共有 n 个样本,则多目标回归任务为学习一个映射函数 $h:\boldsymbol{X}\rightarrow\boldsymbol{Y}$,使得对任意未知的输入向量 $\boldsymbol{x}$,可以同时预测它的所有的输出变量 $\hat{\boldsymbol{y}}=h(\boldsymbol{x})$.

3.2　SI-LSF 算法描述

3.2.1　标签特定特征

为了对目标间相关性进行建模,探索出每个目标的有效特征,避免冗余特征造成的负面影响,本文提出了基于标签特定特征的多目标回归算法(Multi-Target Regression via Label-Specific features,LSF). LSF 可以挖掘出目标间相关性,同时探索与目标最相关的隐含特征信息,其训练与预测框架如下图 1 所示.

如图 1 所示,LSF 的训练与预测阶段均需要进行两段模型的学习或预测. 在训练阶段,需要为每个目标训练第一阶段模型 $\boldsymbol{h}_d$,计算出第一阶段的预测值$\hat{\boldsymbol{Y}}_d$,然后将目标预测值作为附加特征扩展到原始的特征空间 $\boldsymbol{X}$,令变换后的训练集为 $D'=\{(\boldsymbol{x}^{(j)}\cup\hat{\boldsymbol{y}}^{(j)},\boldsymbol{y}^{(j)})\mid 1\leqslant j\leqslant n\}$,其中 $\hat{\boldsymbol{y}}^{(j)}=[\hat{y}_1{}^{(j)},\cdots,\hat{y}_d{}^{(j)}]$表示第一阶段样本的目标预测值,$\boldsymbol{x}^{(j)}$是原始的特征向量,$\boldsymbol{y}^{(j)}$是目标实际值向量,最后训练得到第二阶段的模型. 同理,在预测过程中,先利用$\hat{\boldsymbol{Y}}_d$得到第一阶段预测值,其次学习目标的特定特征,最终用第二阶段模型进行预测.

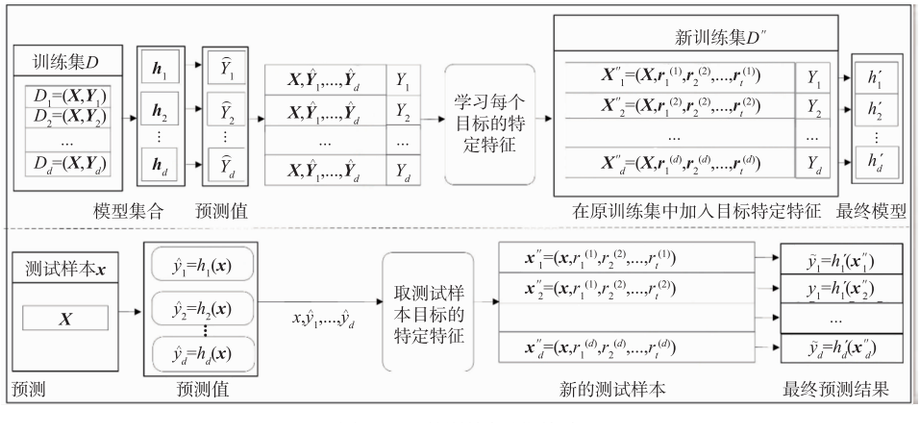


图1　LSF的训练与预测框架

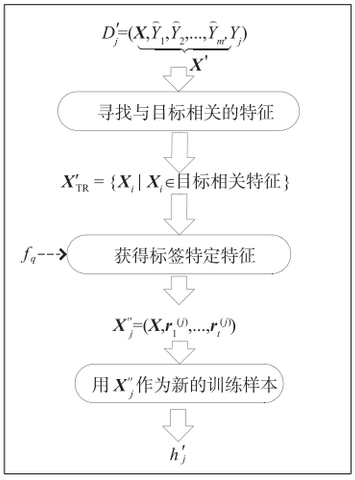


图2 学习标签特定特征框架图

标签特定特征的详细学习过程如上图2所示. 为每个目标 $\boldsymbol{Y}_j$ 寻找与其相关的特征集合 $\boldsymbol{X}'_{\mathrm{TR}}$,利用最小化平方误差找出与 $\boldsymbol{Y}_j$ 相关的特征集合,如式(1)~(3)所示.

$$R_1(\gamma,s)=\{X_i \mid x_{i\gamma}\leqslant s\},R_2(\gamma,s)=\{X_i \mid x_{i\gamma}>s\} \tag{1}$$

$$\hat{c}_b=\frac{\sum_{X_i\in R_1(\gamma,s)} y_{ij}}{|R_b|} \tag{2}$$

$$\min_{\gamma,s}\Big(\min\sum_{X_i\in R_1(\gamma,s)}(y_{ij}-\hat{c}_1)^2+\min\sum_{X_i\in R_2(\gamma,s)}(y_{ij}-\hat{c}_2)^2\Big) \tag{3}$$

式(1)中,s 属于 X_i 中的某个值,利用 s 可以将 X_i 这列属性划分为两部分 $R_1(\gamma,s)$ 和,$R_2(\gamma,s)$,其中 $1\leqslant\gamma\leqslant|X_i|$. 其次对划分好的两部分,利用式(2)计算出对应目标的均值,其中 $b=1,2$. 最后,在式(3)中利用最小化平方误差,找出最好的划分特征及其分裂点 s,从而找出目标的相关特征集合 $\boldsymbol{X}'_{\mathrm{TR}}$. 代码如算法1所示.

算法1 寻找目标相关特征伪代码

输入:训练集 D'_j;目标相关特征数 p
输出:目标相关特征 $\boldsymbol{X}'_{\mathrm{TR}}$
1:令 $\boldsymbol{X}'=(X_1,X_2,\cdots,X_m,\hat{Y}_1,\cdots,\hat{Y}_d)$,$\boldsymbol{X}_{\mathrm{TR}}$ 为空集
2:for $t\leftarrow 1$ to p do
3: for each $X\in\boldsymbol{X}'$ do
4: 利用式(3)计算特征 γ 及其对应分裂点 s 对应的平方误差
5: end for
6: 找出最小平方误差对应的特征 X,γ 及 s,且令 $X_{TR}\leftarrow X$
7: $\boldsymbol{X}'_{\mathrm{TR}}=\boldsymbol{X}'_{\mathrm{TR}}\cup X_{\mathrm{TR}}$
8: 更新 $R_1(\gamma,s)$ 和 $R_2(\gamma,s)$
9:end for

通过算法1选定的目标相关特征,然而,这些相关特征仅是训练集 D'_j 的特征子集,在可能无法充分表达目标. 本文基于目标相关特征构建标签特定特征,增加模型的表达能力,方法如式(4)~(6)所示.

$$g_0(\boldsymbol{X}'_{\mathrm{TR}_0})=f_q(\boldsymbol{X}'_{\mathrm{TR}_0};Y_j) \tag{4}$$

$$r_t^{(j)}=-\left[\frac{\delta l(Y_j,g_{t-1}(\boldsymbol{X}'_{\mathrm{TR}_{t-1}}))}{\delta g_{t-1}(\boldsymbol{X}'_{\mathrm{TR}_{t-1}})}\right] \tag{5}$$

$$g_t(\boldsymbol{X}'_{\mathrm{TR}_t})=g_{t-1}(\boldsymbol{X}'_{\mathrm{TR}_{t-1}})+f_q(\boldsymbol{X}'_{\mathrm{TR}_t};r_t^{(j)}) \tag{6}$$

式(4)是设置初始值,其中 f_q 是基学习器,可以是线性回归,也可以是回归树,则 $g_0(\boldsymbol{X}'_{\mathrm{TR}_0})$ 表示为采用基学习器在输入空间 $\boldsymbol{X}'_{\mathrm{TR}_0}$ 中对目标 Y_j 进行学习的预测值. 式(5)是求当前模型的负梯度值,并将这个负梯度值作为残差估计,其中 $l(\cdot)$ 是平方误差损失函数 $l(Y_j,g_{t-1}(\boldsymbol{X}'_{\mathrm{TR}_{t-1}}))=\frac{1}{2}(Y_j-g_{t-1}(\boldsymbol{X}'_{\mathrm{TR}_{t-1}}))^2$. 实际上,通过求导变换,式(5)等价于 $r_t^{(j)}=Y_j-g_{t-1}(\boldsymbol{X}'_{\mathrm{TR}_{t-1}})$,也就是残差. 本文将这个残差值作为对应目标的标签特定特征. 式(6)是将式(5)中的残差估计作为目标,利用负梯度值对模型进行更新,并作为下一轮迭代的目标. 学习标签特定特征的伪代码如算法2所示.

算法2 学习标签特定特征伪代码

输入:训练集 D'_j;回归模型 f_q;标签特定数 t
输出:标签相关特征 $\boldsymbol{r}_1^{(j)},\cdots,\boldsymbol{r}_t^{(j)}$
1:利用算法3.4在训练集 D'_j 上初始化 $\boldsymbol{X}'_{\mathrm{TR}_0}$
2:令 $g_0(\boldsymbol{X}'_{\mathrm{TR}_0})=f_q(\boldsymbol{X}'_{\mathrm{TR}_0};Y_j)$
3:for $i\leftarrow 1$ to t do
4: 通过公式(5)计算 $r_t^{(j)}$
5: 利用 $f_q(\boldsymbol{X}'_{\mathrm{TR}_t};r_t^{(j)})$,通过式(6)更新 $g_t(\boldsymbol{X}'_{\mathrm{TR}_t})$
6:end for

3.2.2 稀疏集成

LSF能有效挖掘出目标间关联. 然而,LSF在整个训练与预测的流程中,选择的是传统的单目标回归方法,这导致在第一阶段或是第二阶段中无法对复杂的输入与输出关系进行建模. 针对这一问题,本文在LSF上进行改进,提出基于标签特定特征的多目标回归稀疏集成算法(SI-LSF),引入集成学习,增加处理复杂关系的灵活性.

本文提出一种稀疏性的聚合函数:

$$\boldsymbol{w}_j^*=\min_{w_j}\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}_j^{\mathrm{T}}\hat{\boldsymbol{y}}_i-y_{ij})^2+\lambda\|\boldsymbol{w}_j\|_1 \tag{7}$$

$$h_j=\sum_{i=1}^{k}w_{ij}^*f_i \tag{8}$$

式(7)中 $\boldsymbol{w}_j$ 是对基模型分配的权重向量,通过引入 L_1 正则项 $\lambda\|\boldsymbol{w}_j\|_1$,使部分基模型的权重 w_j 为0,从而达到稀疏的目的. 最终的预测结果通过式(8)进行集成计算. 集成过程如图3所示,其中 f_k 表示回归方法,w_{dk} 表示回归方法 f_k 在预测目标 $\boldsymbol{Y}_d$ 时分配的权重,w_{dk} 通过式(7)求得,最后通过求和函数,得到最终的预测结果. 稀疏集成的伪代码如算法3所示.

SI-LSF是在LSF的第一阶段对各个模型的预测结

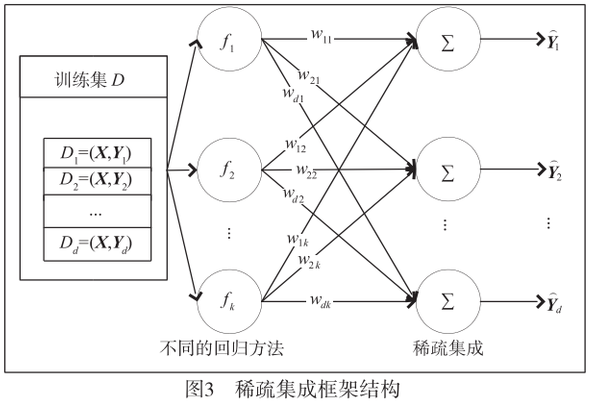


图3　稀疏集成框架结构

果进行集成,优点在于既可使第二阶段标签特定特征更加准确. SI-LSF 的框架如图 4 所示,与 LSF 的框架相同点在于,在训练与预测阶段同样需要进行两段模型的学习或预测,不同之处是无论是训练还是预测,在第一阶段需要对训练集或测试样本进行稀疏集成,从而得到第一阶段模型或预测值.

在考虑了集成学习的聚合函数后,为了处理复杂的输入与输出关系,为了确保稀疏积分的多样性,本文选择不同的回归模型进行集成,可以分为以下三类:线性模型如线性回归(Linear Regression)[17];非线性模型如支持向量回归(Support Vector Regression,SVR)[18];树模型如随机森林(Random Forest,RF)[19].

算法 3　稀疏集成伪代码

输入:训练集 D;回归模型 $f_1\cdots,f_k$;
输出:稀疏集成模型 $h_j,j=1,\cdots,d$
1:for $j=1$ to d do
2:　$D=\{(\boldsymbol{x}^{(1)},\boldsymbol{y}^{(1)}),\cdots,(\boldsymbol{x}^{(n)},\boldsymbol{y}^{(n)})\}$
3:　for $i=1$ to k do
4:　　利用回归方法 f_i 在 D_j 中训练目标 Y_j
5:　end for
6:　计算 $\boldsymbol{w}_j^*=\min\limits_{w_j}\frac{1}{2}\sum\limits_{i=1}^{N}(\boldsymbol{w}_j^{\mathrm{T}}\hat{\boldsymbol{y}}_i-y_{ij})^2+\lambda\|\boldsymbol{w}_j\|_1$
7:　得到最终模型 $h_j=\sum\limits_{i=1}^{k}w_{ij}^*f_i$
8:end for

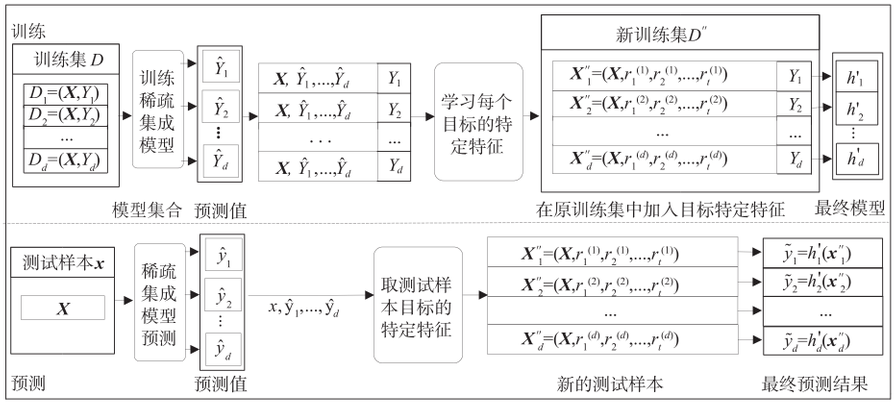


图4　SI-LSF框架图

表 1　数据集信息表

数据集	样本数	属性数	标签数	数据集	样本数	属性数	标签数
andro	49	30	6	osales	639	401	12
slump	103	7	3	enb	768	8	2
edm	154	16	2	wq	1060	16	14
atp7d	296	411	6	sf2	1066	10	3
sf1	323	10	3	scpf	1137	23	3
oes97	334	263	16	scm20d	8966	61	16
atp1d	337	411	6	rf1	9125	64	8
jura	359	15	3	rf2	9125	576	8
oes10	403	298	16	scm1d	9803	280	16

4　实验结果与分析

4.1　数据集与对比方法

本文使用 18 个常用的公开多目标回归数据集进行实验,其详细信息如表 1 所示,其中包含样本数、特征数以及数据集中目标个数. 所有的数据集均来自于 Mulan①. 本文选用 2016 年至 2018 年发表的多目标回归方法进行对比实验,对比方法的详细信息如表 2 所示.

① http://mulan.sourceforge.net/datasets-mlc.html

表 2 实验对比方法信息表

缩写	方法	来源刊物
SST	Stacked SingleTarget[8]	Machine Learning (2016)
ERC	Ensemble of RegressionChains[8]	Machine Learning (2016)
SVRCC	SVR-correlationChains[36]	Information Sciences (2017)
MMR	Multi-layer Multi-targetRegression[15]	PAMI (2018)

4.2 评估指标

在本节中,对所有回归目标,使用平均相对均方根误差 aRRMSE 进行度量,其中 RRMSE 的定义如下:

$$\mathrm{RRMSE}(h,D_{test}) = \sqrt{\frac{\sum_{(x,y)\in D_{test}}(\hat{y}_j - y_j)^2}{\sum_{(x,y)\in D_{test}}(\bar{Y}_j - y_j)^2}} \tag{9}$$

其中 D_{test} 是测试集,$(\boldsymbol{x},\boldsymbol{y})$ 是测试样本,$\boldsymbol{y}$ 是测试集目标的真实数据向量,y_j 是向量 $\boldsymbol{y}$ 中的第 j 个目标的真实值,$\hat{y}_j$ 是通过模型预测得到的第 j 个目标的预测值,$\bar{Y}_j$ 是训练集第 j 个目标的均值. 在测试集中,采用 f 折交叉验证,对每个目标估计 RRMSE 值,即在每一折中对每个目标获得 RRMSE 值,则最终的 RRMSE 值是这个目标所有 RRMSE 的平均值. 为了和其它方法进行公平的比较,与其它方法的实验设置一样,除了在 rf1、rf2 数据集上进行 2 折交叉验证以及在 scm1d、scm20d 数据集上进行 5 折交叉验证,其余数据集均进行 10 折交叉验证[4,16,20]. 通过用不同随机种子数重复交叉验证对比方法的排名稳定性,证明了在这 4 个样本数超 8000 的数据集上 2 折交叉和 5 折交叉的可靠性[5].

由于 RRMSE 只能对数据集上的每个目标进行单一性能测量,不能直接应用于多目标整体的评估. 和大多数研究方法一样,本文将数据集中所有目标变量的平均 RRMSE(即 aRRMSE)作为性能测量.

4.3 实验结果与分析

本文主要从预测性能、标签特定特征的有效性以及稀疏集成的有效性这三方面对算法进行评估与分析,并对算法的关键参数进行分析.

(1)验证不同数据集下算法的预测准确度

SI-LSF 与对比算法 SST,ERC,SVRCC 和 MMR 的多目标预测的具体结果在表 3 中,其中最佳结果用粗体展示,用 AveRank 表示算法的平均排名. 为了验证 SI-LSF 和比较算法之间的在统计上的显著性差异,对表 3 中展示的预测性能使用 Friedman 检验[21],提出原假设 H_0:这 5 种多目标回归算法是等价的,没有显著性差异. 表 4 中总结了 Friedman 检验的统计信息,其中数据集 18 个,对比算法 5 个,显著性水平为 0.05. 正如表 4 中展示的结果,由于 p-value 值远远小于 0.05,拒绝原假设 H_0,在统计上说明了 SI-LSF 和比较算法之间在预测性能上存在显著性差异.

表 3 五个对比算法在不同数据集上的 aRRMSE 值

算法	SST	ERC	SVRCC	MMR	SI-LSF
andro	0.579	0.567	0.446	0.527	**0.357**
atp1d	0.372	0.372	0.378	**0.332**	0.370
atp7d	0.507	0.512	0.534	0.443	**0.428**
edm	0.740	0.741	0.698	0.716	**0.566**
enb	0.121	0.114	0.090	0.111	**0.070**
jura	0.591	0.590	0.589	0.582	**0.545**
oes10	0.421	0.420	0.354	0.403	**0.346**
oes97	0.524	0.524	0.464	0.497	**0.450**
osales	0.726	0.713	0.781	0.709	**0.648**
rf1	0.094	0.091	0.085	0.089	**0.067**
rf2	0.097	0.095	0.086	0.095	**0.058**
scm1d	0.336	0.330	0.324	0.318	**0.312**
scm20d	0.413	0.394	0.386	0.389	**0.347**
scpf	0.831	0.830	0.801	0.812	**0.796**
sf1	1.068	1.089	0.932	0.958	**0.885**
sf2	1.055	1.088	1.030	0.984	**0.886**
slump	0.695	0.689	0.556	0.587	**0.552**
wq	0.909	0.906	0.905	**0.889**	0.898
AveRank	4.500	4..000	2.722	2.500	**1.111**

表 4 Friedman 检验结果表

Friedman 检验	
Friedman 检验值	54.17
p-value($\alpha=0.05$)	4.84e-10
接受或拒绝	拒绝

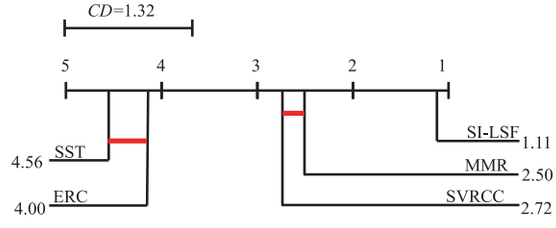


图5 五个对比算法的平均排名(Bonferroni-Dunn检验, $q\alpha$=2.498, α=0.05, k=5, N=18)

在对这 5 种对比算法的整体差异性进行分析比较后,使用 Bonferroni-Dunn 检验作为事后多重检验 post-hoc test[21]. 首先计算出临界差异(Critical Difference, CD)[21],然后以图形方式进行展示. 临界差异是指两种方法被认为显著不同所需的平均排名的最小差异值,它计算方式如下:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{10}$$

其中 k 是比较算法的数量,N 是数据集的数量. 对 Bon-

ferroni-Dunn test 查表可知,在显著性水平 $\alpha=0.05$ 时得到 $q_\alpha=2.498$,因此 $CD=1.32(k=5,N=18)$.

上图 5 中展示了 *CD* 值以及对比算法的平均排名,从图 5 中观察到,SI-LSF 位于图中的最右侧,并且平均排名的差异性大于 *CD* 值,说明 SI-LSF 明显优于 SST,ERC,SVRCC 和 MMR.

(2)验证标签特定特征的有效性

将标签特定特征作为控制变量,把 SI-LSF 与 SI 的预测结果进行比较,方法 SI 遵循 SI-LSF 的相同过程但不使用标签特定特征. 表 5 以 aRRMSE 的形式展示了 SI 和 SI-LSF 的预测结果,其最佳结果用粗体展示.

从表 5 中观察得到,在大部分数据集上 SI-LSF 的预测性能都优于 SI. 利用统计检验的 Wilcoxon signed rank test[21] 在显著性水平 $\alpha=0.05$ 时来判断 SI- LSF 是否明显优于 SI,同时提出原假设:SI 与 SI-LSF 的均值等价,无显著性差异. Wilcoxon signed rank 检验的结果显示在表 6 中. 从表 6 中可以发现由于 $p<0.05$,说明拒绝原假设,也就说明这两种算法的均值不相等,即 SI-LSF 优于 SI,证明:标签特定特征可以明显提高算法的预测准确度.

表 5　SI 和 SI-LSF 算法在不同数据集上的 aRRMSE 值

算法	SI	SI-LSF
andro	0.412	**0.357**
atp1d	0.386	**0.370**
atp7d	0.446	**0.428**
edm	0.593	**0.566**
enb	0.077	**0.070**
jura	0.546	**0.545**
oes10	0.348	**0.346**
oes97	0.450	**0.450**
osales	0.649	**0.648**
rf1	**0.062**	0.067
rf2	0.063	**0.058**
scm1d	0.315	**0.312**
scm20d	0.348	**0.347**
scpf	0.822	**0.796**
sf1	0.902	**0.885**
sf2	0.907	**0.886**
slump	0.561	**0.552**
wq	0.899	**0.898**

表 6　Wilcoxon signed ranks 检验结果

Wilcoxon signed ranks 检验	
对比算法	SI-LSF vs SI
显著性水平 α	0.05
p-value	1.165e-3
结果	胜利

(3)验证稀疏集成的有效性

在 SI-LSF 中,提出稀疏集成处理输入特征与输出目标间的复杂关系. 比较 SI-LSF 与没有稀疏集成学习但其余过程都与 SI-LSF 一致的三种基础回归模型(支持向量回归 SVR-LSF,线性回归 Linear -LSF,随机森林 RF-LSF). 在 SVR-LSF,Linear-LSF 和 RF-LSF 中,SVR,线性回归和 RF 分别用作所有目标的预测模型. SVR-LSF,Linear-LSF,RF-LSF,SI-LSF 在不同数据集上的 aRRMSE 值展示在表 7 中,其中每个数据集的最佳结果也用粗体标记,用 AveRank 表示算法的平均排名.

为了进一步检验 SI-LSF 和 SVR-LSF,Linear-LSF,RF-LSF 之间的性能差异,应用 Friedman 检验这 4 种算法有无显著性差异,提出原假设:这 4 种多目标回归算法是等价的,没有显著性差异. Friedman 检验结果在表 8 中. 从表 8 中观察得到,由于 p-value 远远小于 0.05,因此拒绝原假设,证明这 4 种算法存在差异性. 使用 Bonferroni-Dunn 检验作为事后多重检验 post-hoc test. 计算出临界差异 $CD=1.03$,各对比算法的平均排名如图 6 所示.

表 7　四个对比算法在不同数据集上的 aRRMSE 值

算法	SVR-LSF	Linear-LSF	RF-LSF	SI-LSF
andro	0.415	0.681	0.450	**0.357**
atp1d	0.415	0.418	0.396	**0.370**
atp7d	0.585	0.565	**0.428**	**0.428**
edm	0.611	0.630	0.605	**0.566**
enb	0.472	0.110	0.071	**0.070**
Jura	0.683	0.595	0.584	**0.545**
oes10	0.351	0.869	0.383	**0.346**
oes97	0.454	1.295	0.479	**0.450**
osales	0.834	0.747	0.650	**0.648**
rf1	0.262	0.410	0.073	**0.067**
rf2	0.630	0.372	0.067	**0.058**
scm1d	0.771	0.410	0.313	**0.312**
scm20d	0.348	0.642	0.360	**0.347**
scpf	0.856	0.796	0.834	**0.796**
sf1	0.917	1.229	1.056	**0.885**
sf2	0.930	1.140	1.361	**0.886**
slump	0.604	0.652	0.752	**0.552**
wq	0.939	0.956	**0.898**	**0.898**
AveRank	3.000	3.389	2.444	**1.000**

表 8　Friedman 检验结果表

Friedman 检验	
Friedman 检验值	34.02
p-value($\alpha=0.05$)	1.96e-07
接受或拒绝	拒绝

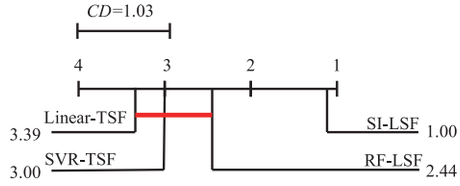


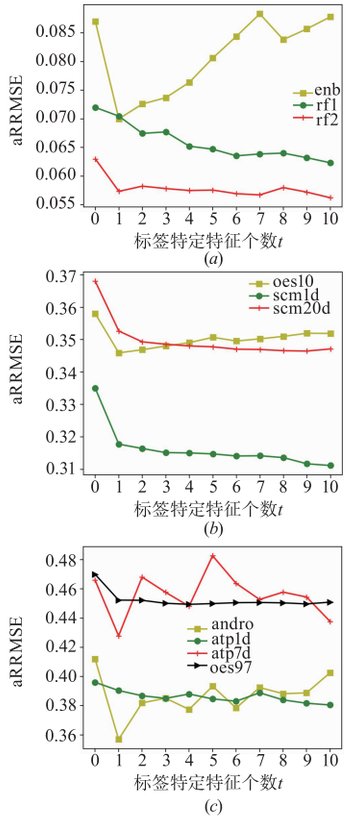
图6 四个对比算法的平均排名(Bonferroni-Dunn检验, α=0.05, k=4, N=18)

从图6中可以看出,使用稀疏集成(即SI-LSF)的方法明显优于那些没有稀疏集成的方法(即SVR-LSF,Linear-LSF和RF-LSF),从而证明了稀疏集成策略的有效性.

(4)参数设置

对于算法SI-LSF,标签特定特征的个数t是一个关键参数.为了研究算法SI-LSF对参数t的敏感度,以1为步长,在范围从0到10内对算法SI-LSF进行实验.如图7所示,图中展示了SI-LSF在不同数据集中不同t值下的aRRMSE值.

从图7中可以发现,当$t=0$,即不加入标签特定特征时,大部分数据集的预测性能都未能到达最佳.在样本数较多且特征数较多的条件下,平均相对误差aRRMSE会随着t的增加而逐渐下降,最后趋于平稳.但是部分数据集,例如图7(a)中的enb,图7(c)中的andro,atp7d,图7(d)中的slump以及图7(e)中的sf2等等会随着t的增加,性能逐渐不稳定,甚至会变差,可能



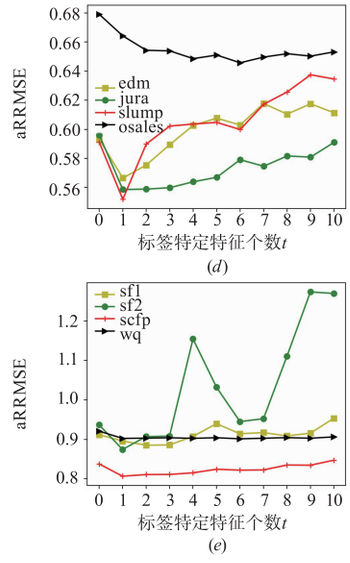


图7 算法SI-LSF在各个数据集下指标aRRMSE随标签特定特征个数t值变化图

的原因是其样本数过少,或是特征数过少,过多的标签特定特征会导致模型过拟合,从而导致预测性能变差.故本文将t值取为1.

5 总结

本文提出了一种基于标签特定特征的多目标稀疏集成方法SI-LSF,通过为单一目标构建标签特定特征来挖掘目标间的相关性.同时利用稀疏性聚合函数对各种回归方法进行集成,从而处理输入与输出的复杂关系,有效地提高了算法的预测性能和多目标处理的灵活性.

虽然标签特定特征可以改善预测性能,然而它们是从与给定目标变量相关的特征子集中学习得到,其不考虑多个目标变量之间的共享信息.下一步将研究如何同时学习每个目标变量和共享共同特征的特定特征,从而提高多目标回归的预测性能.

参考文献

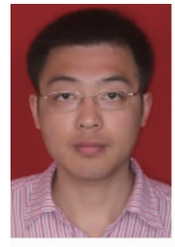
[1] Martin B, Dragi K, Saso D. Ensembles for multi-target regression with random output selections[J]. Machine Learning, 2018, 107(11): 1673 – 1709.

[2] Hadavandi E, Shahrabi J, Shamshirband S. A novel boosted-neural network ensemble for modeling multi-target regression problems[J]. Engineering Applications of Artificial Intelligence, 2015, 45(1): 204 – 219.

[3] Yan Y, Ricci E, Subramanian R, et al. Amulti-task learning framework for head pose estimation under target motion [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(6): 1070 – 1083.

[4] 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.
[5] Borchani H, Varando G, Bielza C, et al. A survey on multi-output regression[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2015, 5(5): 216-233.
[6] Tsoumakas G, Spyromitros-Xioufis E, Vrekou A, et al. Multi-target regression via random linear target combinations[A]. Proceedings of the 2014 Machine Learning and Knowledge Discovery in Databases[C]. Nancy: Springer, 2014. 225-240.
[7] Borchani H, Varando G, Bielza C, et al. A survey on multi-output regression[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2015, 5(5): 216-233.
[8] Wang X, Bi J, Yu S, et al. Multiplicativemultitask feature learning[J]. Journal of Machine Learning Research, 2016, 17(1): 2820-2852.
[9] Chen J, Liu J, Ye J. Learning incoherent sparse and low-rank patterns from multiple tasks[J]. ACM Transactions on Knowledge Discovery from Data, 2012, 5(4): 1-31.
[10] Zhou Q, Zhao Q. Flexibleclustered multi-task learning by learning representative tasks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(2): 266-278.
[11] 张晶. 基于 AdaBoost 回归树的多目标预测算法的研究[D]. 北京:北京交通大学, 2017.
[12] Rothman A J, Levina E, Zhu J. Sparsemultivariate regression with covariance estimation[J]. Journal of Computational and Graphical Statistics, 2010, 19(4): 947-962.
[13] Zhang Y, Yeung D Y. Aconvex formulation for learning task relationships in multi-task learning[A]. Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence[C]. Catalina Island: AUAI press, 2010. 733-742.
[14] Liu H, Wang L, Zhao T. Calibratedmultivariate regression with application to neural semantic basis discovery[J]. Journal of Machine Learning Research, 2015, 16(1): 1579-1606.
[15] Alvarez, Mauricio A, Rosasco L, et al. Kernels for vector-valued functions: a review[J]. Foundations and Trends in Machine Learning, 2011, 4(3): 195-266.
[16] Jacob L, Bach F, Vert J P. Clustered multi-task learning: A convex formulation[A]. Proceedings of the 22nd Annual Conference on Neural Information Processing Systems[C]. Vancouver: Curran Associates, 2008. 745-752.
[17] Yu C, Yao W X. Robust linear regression: a review and comparison[J]. Communications in Statistics-Simulation and Computation, 2017, 46(8): 6261-6282.
[18] Chang C C, Lin C J. A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.
[19] Cutler A, Cutler D R, Stevens J R. Randomforests[J]. Machine Learning, 2004, 45(1): 157-176.
[20] Melki G, Cano A, Kecman V, et al. Multi-target support vector regression via correlation regressor chains[J]. Information Sciences, 2017, 415-416(1): 53-69.
[21] Demšar J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006, 7(1): 1-30.

作者简介



刘洪涛　男,1974 年 9 月生于重庆市北碚,博士,副教授,主要研究方向为社会网络、网络舆情演化、机器学习.
E-mail:liuht@cqupt.edu.cn

李　航　女,1995 年生,硕士研究生,主要研究方向为机器学习与数据挖掘.

王　进　男,1979 年生,工学博士,教授,主要研究方向行业大数据分析及系统架构、大规模数据挖掘与机器学习等.

李鸽鸽　女,1995 年生,硕士研究生,主要研究方向是机器学习与社会网络.