结合目标特定特征和目标相关性的多目标回归

王 进,高选人,张 睿,孙开伟,邓 欣
(重庆邮电大学数据工程与可视计算重点实验室,重庆 400065)

摘 要: 多目标回归旨在使用一组共同的输入变量来预测多个连续变量,其现有方法可归类为问题转换法和算法适应法.它的主要挑战在于如何对输入与输出空间的复杂关系进行建模,以及如何有效利用目标间的相关性.然而,现有的问题转换法很少同时考虑到这两方面.基于此,本文构建了一种问题转换法同时应对这两大挑战,提出了一种结合目标特定特征和目标相关性的多目标回归方法(Multi-Target Regression via Specific Features and Inter-Target Correlations,TSF-TC).TSF-TC 通过对分箱后的样本进行聚类分析构建目标特定特征从而对输入与输出空间的复杂关系进行建模,通过有选择性地堆叠单目标预测值揭示目标间的相关性.本文使用 TSF-TC 在 18 个多目标回归数据集上与现有多目标回归方法进行了对比实验,实验结果充分表明了 TSF-TC 的优势.

关键词: 机器学习;多目标回归;目标特定特征;目标间相关性
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2020)11-2092-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.11.002

Multi-Target Regression via Specific Features and Inter-Target Correlations

WANG Jin,GAO Xuan-ren,ZHANG Rui,SUN Kai-wei,DENG Xin
(*Key Laboratory of Data Engineering and Visual Computing,Chongqing University of Posts and Telecommunications,Chongqing* 400065,*China*)

Abstract: Multi-target regression (MTR) aimed to predict multiple continuous variables using a common set of input variables,whose existing methods could be classified as problem transformation methods (PTM) and algorithm adaptation methods (AAM). Its main challenges were how to model the complex relationship between input and output space,and how to effectively utilize the correlation between targets. However,the existing PTM rarely took both aspects into consideration. So,this paper constructs a problem transformation method named TSF-TC combining target-specific features and Inter-Target correlations. TSF-TC constructs specific features to per target by conducting clustering analysis on the samples after binning, and then reveals the correlation between targets by selectively stacking single target prediction. Comparative experiments with existing multi-target regression methods on 18 datasets fully demonstrate the advantages of TSF-TC.

Key words: machine learning;multi-target regression;target-specific features;inter-target correlations

1 引言

多目标回归(Multi-Target Regression,MTR)是对传统回归模型的扩展,它在同一输入空间中预测多个连续的目标[1].现有的 MTR 方法可分为两大类[2]:(1)问题转换法(Problem Transformation Methods,PTM),将多目标回归问题转换为各个单目标回归问题,每个问题使用传统的回归模型求解;(2)算法适应法(Algorithm Adaptation Methods,AAM),适应特定的单目标回归方法(如决策树),直接处理多目标回归问题.本文主要是基于 PTM 解决 MTR 问题.

目前 MTR 问题主要有以下两大挑战:(1)如何对输入与输出之间的复杂关系建模[3];(2)如何有效利用输出之间的相关性[4].针对第一大挑战,可以为每个目标构建特定特征,即让每个输出对应一套不同的输入;针对第二大挑战,可以使用改进的单目标堆叠算法[1]

收稿日期:2019-12-09;修回日期:2020-06-27;责任编辑:马兰英
基金项目:国家自然科学基金青年科学基金(No.61806033);重庆市自然科学基金面上项目(No.cstc2019jcyj-msxmX0021)

(Stacked Single-Target,SST)揭示目标间的相关性.

现有的 PTM 方法很少同时考虑以上两大问题[5].基于此,本文构建了一种问题转换法同时应对这两大挑战,提出了一种结合目标特定特征和目标相关性的多目标回归方法(Multi-Target Regression via Specific Features and Inter-Target Correlations,TSF-TC).TSF-TC 对输入与输出之间复杂关系建模的方式是:首先对每个单一目标进行分箱,再对每个箱中的样本进行聚类分析得到代表性样本点,最后将原始样本与代表性样本点的距离作为该单一目标的特定特征并扩展到原始特征空间中.TSF-TC 利用输出之间相关性的方式是:首先对每个单一目标进行学习得到预测值,再将与当前目标相关性大的预测值扩展到原始特征空间中.通过扩展目标特定特征和目标相关预测值,能够同时解决 MTR 的两大主要挑战.

与现有的 MTR 方法相比,本文提出的 TSF-TC 方法具有以下优点:

(1)通过提取目标特定特征,TSF-TC 可灵活对输入与输出的复杂关系进行建模.

(2)通过扩展目标相关预测值,TSF-TC 有效利用了目标间的相关性.

(3)作为 PTM 方法,TSF-TC 同时解决了 MTR 的两大主要挑战,提升了预测性能.

2　相关工作

目前 MTR 的研究基本围绕其两大主要挑战进行,可划分为 PTM 和 AAM,即问题转换法和算法适应法[2].PTM 为每个目标构建模型,AAM 使用单个模型同时预测所有目标.常见 MTR 方法如表 1 所示.

表 1　常见 MTR 方法

类别	缩写	方法	来源刊物
PTM	ST	Single-Target[6]	Computer Science(2012)
	RLC	Random Linear Target Combinations[7]	ECML-PKDD (2014)
	SST	Stacked Single-target[1]	Machine Learning (2016)
	ERC	Ensemble of Regressor Chains[1]	Machine Learning (2016)
	SVRCC	MT SVR with Max-correlation Chain[8]	Information Sciences(2017)
	MTR-TSF	Multi-target Regression via Target Specific Features[9]	Knowledge-Based Systems(2019)
AAM	MROTS	Multiple-Output Regression[10]	NIPS(2012)
	OKL	Output Kernel Learning[11]	Foundations and Trends in Machine Learning(2011)
	MSLR	Multi-Target Sparse Latent rRgressio[12]	IEEE TNNLS(2018)
	MMR	Multi-Layer Multi-Target Regression[13]	IEEE TPAMI(2018)

常见的 PTM 主要包括单目标方法[6](Single-Target,ST)、随机目标组合[7](Random Linear Target Combinations,RLC)、单目标堆栈[1](Stacked Single-Target,SST)、组合回归链[1](Ensemble of Regressor Chains,ERC)、具有最大相关链的多目标 SVR[8](MT SVR with Max-Correlation Chain,SVRCC)以及基于目标特定特征的多目标回归方法[9](Multi-Target Regression via Target Specific Features,MTR-TSF)等.ST 直接对每个目标基于原始输入构建回归模型[6],没有解决 MTR 的两大挑战;RLC 通过目标间的随机线性组合构造新的目标[7],利用了目标间相关性,复杂度较高;SST 是对 ST 的扩展,包含两层结构,第一层使用原始输入空间构建回归模型,第二层将第一层的目标预测值堆叠到原始输入空间进行第二层的训练预测[1],但堆叠的无关目标的预测值会使模型性能急剧退化;ERC 通过构建多个回归链模型后采用多数投票得到最终结果[1],然而链的选择会对模型性能产生一定影响并且复杂度较高;SVRCC 通过创建最大相关链[8]利用目标间的相关性简化了 ERC 算法,但没能考虑到最大相关链的局部相关性;MTR-TSF[9]通过层次 K-Means 聚类结合 boosting 的策略构造目标相似度矩阵,再基于相似度矩阵通过 K-Medoids 聚类得到目标特定特征,复杂度过高.

常见的 AAM 主要包括多输出回归模型[10](Multiple-Output Regression,MROTS)、输出内核学习[11](Output Kernel Learning,OKL)、多目标稀疏隐空间回归[12](Multi-Target Sparse Latent rRgression,MSLR)以及多层多目标回归[13](Multi-Layer Multi-Target Regression,MMR)等.MROTS 利用了潜在模型参数的协方差矩阵以及观测输出的条件协方差矩阵,专注于探索目标间相关性[10],但对于非线性的输入输出关系没有提供理论支撑.OKL 通过学习多个目标的输出内核来处理复杂的输入输出关系[11],但不能完全挖掘出目标间的相关性.MMR 与 MSLR 都采用相似的方式即核技巧解决复杂的输入输出关系,为了有效利用目标间的相关性,

MSLR 通过稀疏学习[12],MMR 通过矩阵弹性网(Matrix Elastic Nets,MEN)编码[13].

一般认为,AAM 更容易建模和解释目标间的依赖关系[5],而通过提取目标特定特征[14]的 PTM 更能灵活地处理输入与输出间的复杂关系. 本文提出的 TSF-TC 方法基于 PTM 的思想,但同时解决了 MTR 的两大主要挑战.

3 结合目标特定特征和目标相关性的多目标回归

3.1 符号定义

在 MTR 问题中,设 $\boldsymbol{D}=(\boldsymbol{X},\boldsymbol{Y})$ 为 n 个样本的训练集,其中 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$ 是由 d 个特征向量 $(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_j,\cdots,\boldsymbol{X}_d)$ 构成的输入空间,$\boldsymbol{Y}\in\mathbb{R}^{n\times m}$ 是由 m 个目标向量 $(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_i,\cdots,\boldsymbol{Y}_m)$ 构成的输出空间. $(\boldsymbol{x}^{(l)},\boldsymbol{y}^{(l)})$ 表示一个样本,包含 d 维输入变量 $(x_1^{(l)},\cdots,x_j^{(l)},\cdots,x_d^{(l)})$ 和 m 维输出变量 $(y_1^{(l)},\cdots,y_i^{(l)},\cdots,y_m^{(l)})$. 其中 $i\in\{1,\cdots,m\}$,$j\in\{1,\cdots,d\}$,$l\in\{1,\cdots,n\}$. MTR 任务就是学习一个映射函数 $h:\boldsymbol{X}\rightarrow\boldsymbol{Y}$,使得对任意未知的输入向量 $\boldsymbol{x}\in\mathbb{R}^{1\times d}$,可同时预测出其所有的目标 $\hat{\boldsymbol{y}}=h(\boldsymbol{x})$.

3.2 TSF-TC 算法描述

3.2.1 目标特定特征

为了解决 MTR 问题的第一大挑战,即对输入与输出的复杂关系进行建模,需要构建每个目标的特定特征,本节提出了基于目标特定特征的多目标回归算法(Multi-Target Regression via Specific Features,TSF).

TSF 通过对分箱后的样本进行聚类分析构建目标特定特征. TSF 在构建目标特定特征时,受到了多标签分类算法 LIFT[14]的启发. LIFT 分正负样本分别聚类,以原始特征与聚类中心点的距离作为标签特定特征. 而多目标回归的每个目标都是连续值,没有正负样本之分,所以 TSF 首先对每个目标进行分箱.

TSF 提出了一种同时考虑到目标与特征的基于 K-Means 聚类的分箱方式. 具体地,算法对训练集 $\boldsymbol{D}$ 的特征部分 $\boldsymbol{X}$ 进行列归一化操作得到 $\overline{\boldsymbol{X}}$,标签部分 $\boldsymbol{Y}$ 不变,进而得到归一化后的训练集 $\overline{\boldsymbol{D}}=(\overline{\boldsymbol{X}},\boldsymbol{Y})$. 对每个目标以 $\overline{\boldsymbol{D}}_i=(\overline{\boldsymbol{X}},\boldsymbol{Y}_i)$ 作为输入进行 K-Means 聚类从而分箱. 对于每组目标与归一化后的特征,分别令聚类数 K 值为 2~20,计算每个 K 值下的簇内误差平方和 SSE[15]:

$$SSE=\sum_{q=1}^{K}\sum_{\boldsymbol{p}\in\boldsymbol{C}_q}|\boldsymbol{p}-\boldsymbol{m}_q|^2 \tag{1}$$

式中,$\boldsymbol{C}_q$ 为第 q 个簇;$\boldsymbol{p}$ 为 $\boldsymbol{C}_q$ 中的样本点;$\boldsymbol{m}_q$ 为 $\boldsymbol{C}_q$ 的质心即 $\boldsymbol{C}_q$ 中所有样本的均值.

SSE 表示所有样本的聚类误差,代表了聚类效果的好坏.

由此得出了每个目标下 2~20 不同 K 值的 SSE,根据"手肘法"[15]可确定每个目标的最佳聚类数 B_i,即分箱个数,从而得到分箱后的数据集 $\boldsymbol{D}_i=\{(\boldsymbol{X}_{(i)}^1,\boldsymbol{Y}_i^1),\cdots,(\boldsymbol{X}_{(i)}^b,\boldsymbol{Y}_i^b),\cdots,(\boldsymbol{X}_{(i)}^{B_i},\boldsymbol{Y}_i^{B_i})\}$,其中 $b\in\{1,\cdots,B_i\}$,$i\in\{1,\cdots,m\}$,$(\boldsymbol{X}_{(i)}^b,\boldsymbol{Y}_i^b)$ 表示第 i 个目标的第 b 个箱. 分箱算法的伪代码如算法 1 所示.

算法 1 分箱算法

输入:训练集 $\boldsymbol{D}=(\boldsymbol{X},\boldsymbol{Y})$

输出:每个目标分箱后的数据集 $\boldsymbol{D}_i=\{(\boldsymbol{X}_{(i)}^1,\boldsymbol{Y}_i^1),\cdots,(\boldsymbol{X}_{(i)}^b,\boldsymbol{Y}_i^b),\cdots,(\boldsymbol{X}_{(i)}^{B_i},\boldsymbol{Y}_i^{B_i})\}$,$i\in\{1,\cdots,m\}$

```
D 的特征部分 X=(X_1,…,X_j,…,X_d) 按列归一化为 X̄=(X̄_1,…,X̄_j,…,X̄_d),标签部分不变
令归一化的训练集为 D̄=(X̄,Y)
for i←1 to m do
    for k←2 to 20 do
        对第 i 个目标对应的归一化数据集 D̄_i=(X̄,Y_i) 进行 K-Means 聚类,聚 k 类
        利用式(1)计算簇内误差平方和 SSE
    end for
    根据"手肘法"[15]得到第 i 个目标的最佳聚类数 B_i
    归一化数据集 D̄_i=(X̄,Y_i) 进行 K=B_i 的 K-Means 聚类
    得到第 i 个目标分箱后的归一化数据集 D̄_i={(X̄^1_(i),Y_i^1),…,(X̄^b_(i),Y_i^b),…,(X̄^{B_i}_(i),Y_i^{B_i})}
    对应于第 i 个目标分箱后的归一化前的数据集 D_i={(X^1_(i),Y_i^1),…,(X^b_(i),Y_i^b),…,(X^{B_i}_(i),Y_i^{B_i})}
end for
```

分箱完成后 TSF 就可以使用类似多标签分类算法 LIFT[14]的思想寻找每个目标的特定特征. 具体地,对第 i 个目标 $\boldsymbol{Y}_i(1\leqslant i\leqslant m)$,分箱后的数据集为 $\boldsymbol{D}_i=\{(\boldsymbol{X}_{(i)}^1,\boldsymbol{Y}_i^1),\cdots,(\boldsymbol{X}_{(i)}^b,\boldsymbol{Y}_i^b),\cdots,(\boldsymbol{X}_{(i)}^{B_i},\boldsymbol{Y}_i^{B_i})\}$,对第 b 个箱的特征部分 $\boldsymbol{X}_{(i)}^b(1\leqslant i\leqslant m,1\leqslant b\leqslant B_i)$ 使用 K-Means 算法将其划分为 K_i 个簇,其中心点表示为 $\boldsymbol{C}_i^b=\{\boldsymbol{C}_i^{b(1)},\cdots,\boldsymbol{C}_i^{b(k)},\cdots,\boldsymbol{C}_i^{b(K_i)}\}$,其中 $\boldsymbol{C}_i^{b(k)}\in\mathbb{R}^{1\times d}$,$b\in\{1,\cdots,B_i\}$,$i\in\{1,\cdots,m\}$,$k\in\{1,\cdots,K_i\}$. 则对于第 i 个目标,总共可以构造 $B_i\times K_i$ 个中心点 $\boldsymbol{C}_i=\{\boldsymbol{C}_i^1,\cdots,\boldsymbol{C}_i^b,\cdots,\boldsymbol{C}_i^{B_i}\}$.

簇中心 $\boldsymbol{C}_i=\{\boldsymbol{C}_i^1,\cdots,\boldsymbol{C}_i^b,\cdots,\boldsymbol{C}_i^{B_i}\}$ 表征了第 i 个目标输入空间每个箱中的底层结构,每个中心点都是最能代表该目标的样本,可用作构造第 i 个目标特定特征的基础. 具体地,TSF 创建从原始 d 维空间 χ 到 $B_i\times K_i$ 维目标特定空间 Z_i 映射 $\phi_i:\chi\rightarrow Z_i$,如下所示:

$$\phi_i(\boldsymbol{x})=[d(\boldsymbol{x},\boldsymbol{C}_i^{(1)}),\cdots,d(\boldsymbol{x},\boldsymbol{C}_i^{(i')}),\cdots,d(\boldsymbol{x},\boldsymbol{C}_i^{(B_iK_i)})] \tag{2}$$

式中,$d(\cdot,\cdot)$ 为两个实例之间的欧氏距离;$\boldsymbol{C}_i^{(i')}$ 为第 i 个目标的第 i' 个中心点,$\boldsymbol{C}_i^{(i')}\in\mathbb{R}^{1\times d}$,其中 $i\in\{1,\cdots,m\}$,$i'\in\{1,\cdots,B_iK_i\}$.

由此,根据训练集 $\boldsymbol{D}$ 的特征部分 $\boldsymbol{X}$,可得到第 i 个目标的特定特征 $\boldsymbol{SF}_i=\phi_i(\boldsymbol{X})$,且 $\boldsymbol{SF}_i\in\mathbb{R}^{n\times B_iK_i}$.

在这里令第 i 个目标的每个箱都聚为相同的 K_i 个簇,通过这种方式,从每个箱中获得的聚类信息被同等重视. 具体地,聚类数 K_i 设置为:

$$K_i = r \times \min(|\boldsymbol{X}_{(i)}^1|, \cdots, |\boldsymbol{X}_{(i)}^b|, \cdots, |\boldsymbol{X}_{(i)}^{B_i}|) \quad (3)$$

式中,$|\boldsymbol{X}_{(i)}^b|$ 为第 i 个目标第 b 个箱的特征部分,其中 $i \in \{1, \cdots, m\}$,$b \in \{1, \cdots, B_i\}$;r 为控制目标特定特征数目的比率参数,$r \in (0,1]$.

寻找目标特定特征的伪代码如算法 2 所示.

算法 2　寻找目标特定特征

输入:分箱数据集 $\boldsymbol{D}_i = \{(\boldsymbol{X}_{(i)}^1, \boldsymbol{Y}_i^1), \cdots, (\boldsymbol{X}_{(i)}^b, \boldsymbol{Y}_i^b), \cdots, (\boldsymbol{X}_{(i)}^{B_i}, \boldsymbol{Y}_i^{B_i})\}$, $i \in \{1, \cdots, m\}$;比率参数 $r, r \in (0,1]$

输出:原始特征到目标特定特征的映射 ϕ_i

```
for i←1 to m do
    利用式(3)计算聚类数 K_i = r × min(|X_(i)^1|, ⋯, |X_(i)^b|, ⋯, |X_(i)^{B_i}|)
    for b←1 to B_i do
        对特征部分 X_(i)^b 进行 K-Means 聚类,聚 K_i 类,得 K_i 个中心点 C_i^b = {C_i^{b(1)}, ⋯, C_i^{b(k)}, ⋯, C_i^{b(K_i)}}
    end for
    得到 B_i × K_i 个中心点 C_i = {C_i^1, ⋯, C_i^b, ⋯, C_i^{B_i}}
    利用式(2)得到原始特征到目标特定特征的映射 φ_i
end for
```

接着,TSF 可根据原始特征 $\boldsymbol{X}$ 与映射 ϕ_i 得到训练集每个目标的特定特征 $\boldsymbol{SF}_i = \phi_i(\boldsymbol{X})$. TSF 旨在用原始特征 $\boldsymbol{X}$ 与生成的每个目标的特定特征 $\boldsymbol{SF}_i$ 推导出 m 个回归模型 $\{h_1, \cdots, h_i, \cdots, h_m\}$.

第 i 个目标的特定特征 $\boldsymbol{SF}_i$ 是对原始输入空间 $\boldsymbol{X}$ 的扩展. $\boldsymbol{SF}_i$ 经过目标分箱(挖掘每个目标在特征空间的不同分布)、分箱后的特征空间聚类(得到每个目标的代表性样本点)、原始特征与簇中心点求距离(特征组合)得到. 类似推荐算法中经典的因子分解机[16](Factorization Machines, FM),通过特征组合提升了逻辑回归[17](Logistic Regression, LR)的性能,即:

$$\hat{y} = w_0 + \sum_{j=1}^{d} w_j x_j + \sum_{j=1}^{d-1} \sum_{j'=j+1}^{d} \langle \boldsymbol{v}_j, \boldsymbol{v}_{j'} \rangle x_j x_{j'} \quad (4)$$

式中,w_0 为 LR 的偏置项;w_j 为第 j 个特征的权重,$j \in \{1, \cdots, d\}$;x_j 为 $\boldsymbol{x}$ 的第 j 个特征,$j \in \{1, \cdots, d\}$;$\boldsymbol{v}_j$ 为第 j 个目标的隐向量,$\boldsymbol{v}_j \in \mathbb{R}^{1 \times k}$;$k$ 为算法的超参数;$\langle \cdot, \cdot \rangle$ 为向量内积.

式(4)为了书写方便省略了 sigmoid 函数. FM 算法两个特征的隐向量作内积表示特征的组合,有效提升了 LR 算法的性能.

TSF 算法的目标特定特征 $\boldsymbol{SF}_i$ 一定程度上也是对特征的组合,即:

$$\hat{y}_i^{(l)} = w_0 + \sum_{j=1}^{d} w_j x_j^{(l)} + \sum_{i'=1}^{B_i K_i} w_{i'} \boldsymbol{SF}_{ii'}^{(l)} \quad (5)$$

式中,$w_{i'}$ 为 $\boldsymbol{SF}_i$ 第 i' 个特征值的权重,$i \in \{1, \cdots, m\}$,$i' \in \{1, \cdots, B_i K_i\}$;$\boldsymbol{SF}_{ii'}^{(l)}$ 为第 l 个样本的 $\boldsymbol{SF}_i$ 的第 i' 个特征值,$l \in \{1, \cdots, n\}$.

而

$$\boldsymbol{SF}_{ii'}^{(l)} = \sqrt{(x_1^{(l)} - c_{i1}^{(i')})^2 + \cdots + (x_j^{(l)} - c_{ij}^{(i')})^2 + \cdots + (x_d^{(l)} - c_{id}^{(i')})^2},$$

则式(5)可化为:

$$\hat{y}_i^{(l)} = w_0 + \sum_{j=1}^{d} w_j x_j^{(l)} + \sum_{i'=1}^{B_i K_i} \cdot w_{i'} \sqrt{(x_1^{(l)} - c_{i1}^{(i')})^2 + \cdots + (x_j^{(l)} - c_{ij}^{(i')})^2 + \cdots + (x_d^{(l)} - c_{id}^{(i')})^2} \quad (6)$$

式中,$c_{ij}^{(i')}$ 为第 i 个目标的第 i' 个中心点的第 j 个值,$j \in \{1, \cdots, d\}$.

式(6)的第三项表示了特征间的组合,可有效提升单个目标的预测性能. 而每个目标的特定特征 $\boldsymbol{SF}_i$ 是与该目标对应的中心点求距离而得,代表了该目标对应的特征组合. 因此为每个目标提取特定特征并扩展原始输入空间能有效提高多目标回归问题的预测效果.

具体地,对第 i 个目标,通过极端梯度提升树[18](Extreme Gradient Boosting, XGBoost)模型学习一个映射函数 $h_i: \boldsymbol{X} + \boldsymbol{SF}_i \rightarrow \boldsymbol{Y}_i$. 基于 h_i 和 ϕ_i,对任意一个测试样本 $\boldsymbol{x}^{(q)}$,算出其目标特定特征 $\boldsymbol{SF}_i^{(q)} = \phi_i(\boldsymbol{x}^{(q)})$ 后,即可求得 $\boldsymbol{x}^{(q)}$ 第 i 个目标的预测值为 $\hat{y}_i^{(q)} = h_i(\boldsymbol{x}^{(q)} + \boldsymbol{SF}_i^{(q)})$.

TSF 的伪代码如算法 3 所示.

算法 3　TSF

输入:训练集 $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Y})$;原始特征到目标特定特征的映射 $\phi_i, i \in \{1, \cdots, m\}$;测试样本 $\boldsymbol{x}^{(q)}$

输出:测试样本 $\boldsymbol{x}^{(q)}$ 的预测值 $\hat{\boldsymbol{y}}^{(q)} = (\hat{y}_1^{(q)}, \cdots, \hat{y}_i^{(q)}, \cdots, \hat{y}_m^{(q)})$

```
for i←1 to m do
    计算得到训练集的目标特定特征为 SF_i = φ_i(X)
    学习一个映射函数 h_i: X + SF_i → Y_i
end for
for i←1 to m do
    计算得到测试样本的目标特定特征为 SF_i^(q) = φ_i(x^(q))
    根据 h_i 得到测试样本第 i 个目标的预测值为 ŷ_i^(q) = h_i(x^(q) + SF_i^(q))
end for
得到测试样本 x^(q) 的预测值 ŷ^(q) = (ŷ_1^(q), ⋯, ŷ_i^(q), ⋯, ŷ_m^(q))
```

综上,TSF 的训练与测试框架如图 1 所示.

3.2.2　目标相关性

TSF 通过分箱与聚类分析能够有效挖掘出每个目标的特定特征,可灵活对输入与输出空间的复杂关系进行建模. 然而,TSF忽略了目标间的相关性. 由此,本

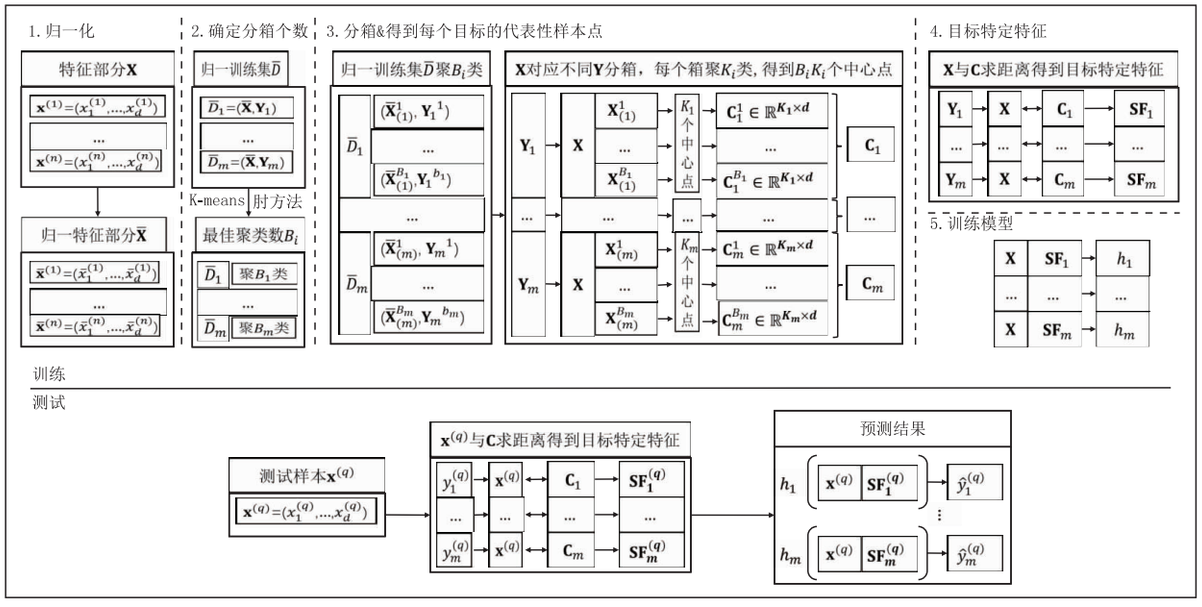

图1 TSF的训练与预测框架

节进一步提出一种结合目标特定特征和目标相关性的多目标回归方法(Multi-Target Regression via Specific Features and Inter-Target Correlations,TSF-TC).

在考虑目标相关性的多目标回归算法中,SST[1]是比较经典的一种,通过堆叠目标预测值从而考虑了目标间的相关性. TSF-TC 受到了 SST 算法的启发,包含了两层训练预测.

TSF-TC 首先计算了目标间的皮尔逊相关系数[19]. 对于第 i 个目标,易得到与目标 $\boldsymbol{Y}_i$ 相关性的绝对值大于阈值 t 的目标索引集合:

$$Y_i^{corr}=\{i'\mid |P(\boldsymbol{Y}_i,\boldsymbol{Y}_{i'})|>t,\ 1\leqslant i\leqslant m-1,i+1\leqslant i'\leqslant m,0\leqslant t<1\} \tag{7}$$

式中,$\boldsymbol{Y}_i$ 为第 i 个目标向量,其中 $i\in\{1,\cdots,m-1\}$;$\boldsymbol{Y}_{i'}$ 为除了第 i 个目标外的其余某个目标向量,其中 $i'\in\{i+1,\cdots,m\}$;$|\cdot|$ 为绝对值;t 为阈值,$t\in[0,1)$.

TSF-TC 的第一层训练预测就是 TSF 算法. 通过分箱与聚类分析分别得到训练集的预测值 $\hat{\boldsymbol{Y}}=(\hat{\boldsymbol{Y}}_1,\cdots,\hat{\boldsymbol{Y}}_i,\cdots,\hat{\boldsymbol{Y}}_m)$ 和测试样本的预测值 $\hat{\boldsymbol{y}}^{(q)}=(\hat{y}_1^{(q)},\cdots,\hat{y}_i^{(q)},\cdots,\hat{y}_m^{(q)})$. 如果是 TSF,直接输出 $\hat{\boldsymbol{y}}^{(q)}$ 作为最终结果. TSF-TC 则包含了一层额外的训练预测.

TSF-TC 的第二层训练预测,对于第 i 个目标,训练集堆叠目标索引集合 Y_i^{corr} 的训练集预测值 $\hat{\boldsymbol{Y}}_i^{corr}$,测试样本堆叠目标索引集合 Y_i^{corr} 的测试样本预测值 $\hat{\boldsymbol{y}}_i^{corr(q)}$,其中:

$$\hat{\boldsymbol{Y}}_i^{corr}=\{\hat{\boldsymbol{Y}}_{i'}\mid i'\in Y_i^{corr}\},\hat{\boldsymbol{y}}_i^{corr(q)}=\{\hat{y}_{i'}^{(q)}\mid i'\in Y_i^{corr}\} \tag{8}$$

式中,Y_i^{corr} 为与第 i 个目标的相关性大于阈值 t 的目标索引集合;$\hat{\boldsymbol{Y}}_{i'}$ 为训练集第 i' 个目标的预测值;$\hat{y}_i^{(q)}$ 为测试样本第 i' 个目标的预测值.

TSF-TC 旨在根据训练集的原始特征 $\boldsymbol{X}$、目标特定特征 $\boldsymbol{SF}_i$ 和第一层相关目标预测值 $\hat{\boldsymbol{Y}}_i^{corr}$ 推导出 m 个回归模型 $\{h_1',\cdots,h_i',\cdots,h_m'\}$. 具体地,对第 i 个目标,通过 XGBoost 模型学习一个映射函数 $h_i':\boldsymbol{X}+\boldsymbol{SF}_i+\hat{\boldsymbol{Y}}_i^{corr}\rightarrow\boldsymbol{Y}_i$. 对于任意一个测试样本 $\boldsymbol{x}^{(q)}$,根据其原始特征 $\boldsymbol{x}^{(q)}$、目标特定特征 $\boldsymbol{SF}_i^{(q)}$ 和第一层相关目标预测值 $\hat{\boldsymbol{y}}_i^{corr(q)}$,可求得第 i 个目标的预测值为 $\hat{\hat{\boldsymbol{y}}}_i^{(q)}=h_i'(\boldsymbol{x}^{(q)}+\boldsymbol{SF}_i^{(q)}+\hat{\boldsymbol{y}}_i^{corr(q)})$. TSF-TC 的伪代码如算法 4 所示.

算法 4 TSF-TC

输入:训练集 $\boldsymbol{D}$;原始特征到目标特定特征映射 $\phi_i,i\in\{1,\cdots,m\}$;阈值 $t,t\in[0,1)$;测试样本 $\boldsymbol{x}^{(q)}$

输出:测试样本 $\boldsymbol{x}^{(q)}$ 的预测值 $\hat{\boldsymbol{y}}^{(q)}=(\hat{y}_1^{(q)},\cdots,\hat{y}_i^{(q)},\cdots,\hat{y}_m^{(q)})$

```
for i←1 to m − 1 do
    for i′←i + 1 to m do
        计算目标 Y_i 与目标 Y_i′ 的皮尔逊相关系数 P(Y_i, Y_i′)
    end for
end for
for i←1 to m do
    利用式(7)得到与目标 Y_i 相关性的绝对值大于阈值 t 的目标索引集合 Y_i^corr
    计算得到训练集的目标特定特征为 SF_i = φ_i(X)
    学习一个映射函数 h_i: X + SF_i → Y_i
    根据 h_i 得到训练集第 i 个目标的预测值 Ŷ_i = h_i(X + SF_i)
end for
for i←1 to m do
    计算得到测试样本的目标特定特征为 SF_i^(q) = φ_i(x^(q))
    根据 h_i 得到测试样本第 i 个目标的预测值为 ŷ_i^(q) = h_i(x^(q) + SF_i^(q))
end for
```

16:for i←1 to m do

17: 利用式(8)得到与第 i 个目标的相关性大于阈值 t 的训练集目标集合的预测值 $\hat{\boldsymbol{Y}}_i^{corr}$

18: 学习一个映射函数 h_i':$\boldsymbol{X}+\boldsymbol{SF}_i+\hat{\boldsymbol{Y}}_i^{corr}\rightarrow\boldsymbol{Y}_i$

19:end for

20:for i←1 to m do

21: 利用式(8)得到与第 i 个目标的相关性大于阈值 t 的测试样本目标集合的预测值 $\hat{y}_i^{corr(q)}$

22: 根据 h_i'得到测试样本第 i 个目标的预测值为 $\hat{\hat{\boldsymbol{y}}}_i^{(q)}=h_i'(\boldsymbol{x}^{(q)}+\boldsymbol{SF}_i^{(q)}+\hat{\boldsymbol{y}}_i^{corr(q)})$

23:end for

综上,TSF-TC 的训练与测试框架如图 2 所示.

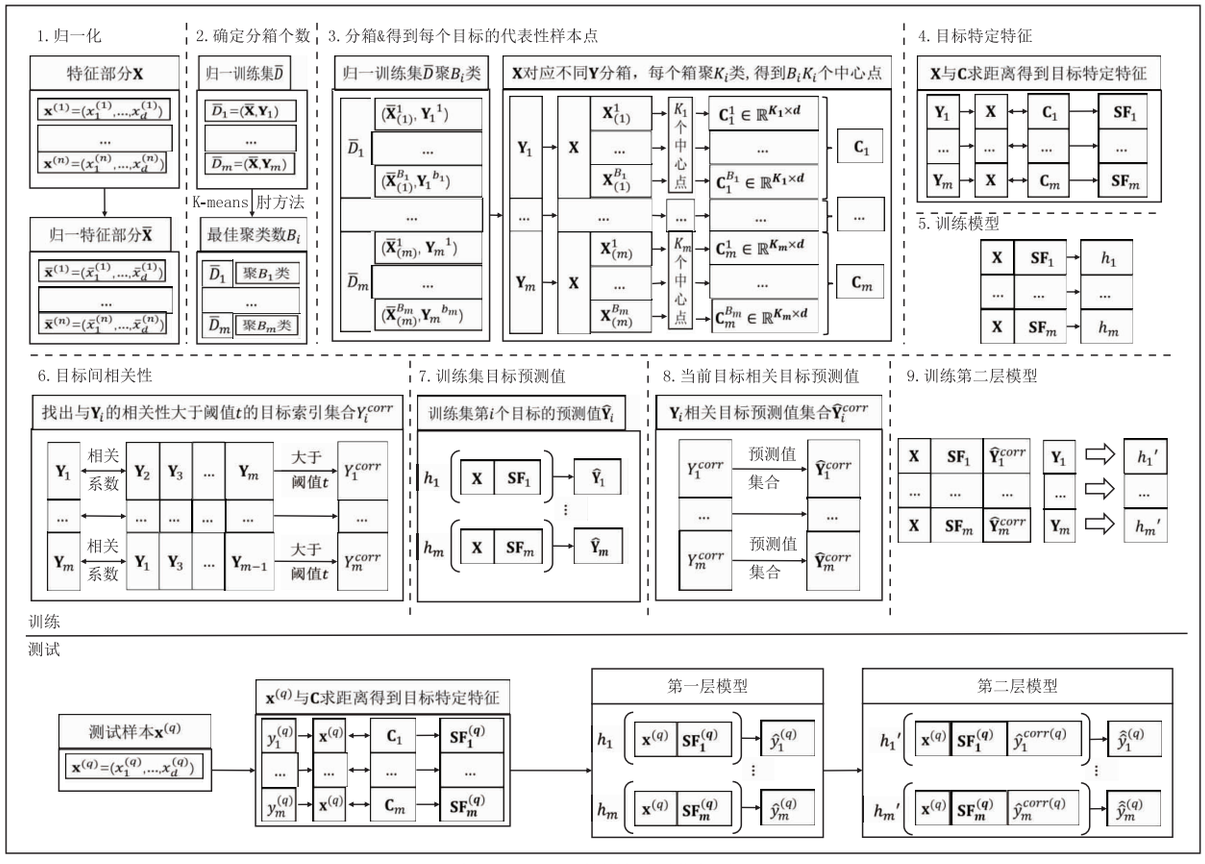

图2　TSF-TC的训练与预测框架

4　实验结果与分析

本节在 18 个数据集上使用 TSF-TC 算法、TSF 算法和 7 个多目标回归算法进行了性能比较并引入假设检验中的 Friedman 检验[20]和 Nemenyi 后续检验[20].

4.1　数据集与对比方法

本节使用了 18 个公开的多目标回归数据集进行实验,其详细信息如表 2 所示,其中包含样本数、特征数和目标数. 所有数据集均来自于 Mulan①.

本节选用 2011 年 ~ 2018 年发表的 7 个多目标回归算法进行对比实验,对比算法的详细信息如表 3 所示.

4.2　评估指标

多目标回归通常使用平均相对均方根误差 aRRMSE[13]进行度量,其中相对均方根误差 RRMSE 的定义如下:

$$\mathrm{RRMSE}(h,\boldsymbol{D}_{\mathrm{test}})=\sqrt{\frac{\sum_{(\boldsymbol{x}^{(q)},\boldsymbol{y}^{(q)})\in\boldsymbol{D}_{\mathrm{test}}}(y_i^{(q)}-\hat{y}_i^{(q)})^2}{\sum_{(\boldsymbol{x}^{(q)},\boldsymbol{y}^{(q)})\in\boldsymbol{D}_{\mathrm{test}}}(y_i^{(q)}-\bar{y}_i)^2}}\tag{9}$$

式中,h 为多目标回归模型;$\boldsymbol{D}_{test}$为测试集;$(\boldsymbol{x}^{(q)},\boldsymbol{y}^{(q)})$为测试样本,$\boldsymbol{x}^{(q)}$表示输入向量,$\boldsymbol{y}^{(q)}$表示目标向量;$y_i^{(q)}$为测试样本第 i 个目标的真实值;$\hat{y}_i^{(q)}$ 为测试样本第 i 个目标的预测值;$\bar{y}_i$ 为训练集 $\boldsymbol{D}$ 的第 i 个目标的均值.

RRMSE(h,$\boldsymbol{D}_{\mathrm{test}}$)的值越小,代表模型 h 的性能越好.

这里采用 k 折交叉验证算出数据集每一折的 RRMSE,最终可求得 k 个 RRMSE 值,aRRMSE 则是这 k 个 RRMSE 的均值:

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

表 2 多目标回归数据集信息表

数据集	样本数	特征数	目标数	数据集	样本数	特征数	目标数
andro	49	30	6	rf1	9125	64	8
atp1d	337	411	6	rf2	9125	576	8
atp7d	296	411	6	scm1d	9803	280	16
edm	154	16	2	scm20d	8966	61	16
enb	768	8	2	scpf	1137	23	3
jura	359	15	3	sf1	323	10	3
oes10	403	298	16	sf2	1066	10	3
oes97	334	263	16	slump	103	7	3
osales	639	413	12	wq	1060	16	14

表 3 实验对比方法信息表

缩写	方法	来源刊物
OKL	Output Kernel Learning[11]	Foundations and Trends in Machine Learning(2011)
MROTS	Multiple-Output Regression[10]	NIPS(2012)
RLC	Random Linear Target Combinations[7]	ECML-PKDD (2014)
SST	Stacked Single-target[1]	Machine Learning (2016)
ERC	Ensemble of Regressor Chains[1]	Machine Learning (2016)
SVRCC	MT SVR with Max-correlation Chain[8]	Information Sciences (2017)
MMR	Multi-Layer Multi-Target Regression[13]	IEEE TPAMI (2018)

$$\mathrm{aRRMSE}(h,\boldsymbol{D}) = \frac{1}{k}\sum_{i=1}^{k}\mathrm{RRMSE}(h^i,\boldsymbol{D}^i) \quad (10)$$

式中,$\boldsymbol{D}$ 为训练集;$\boldsymbol{D}^i$ 为训练集 $\boldsymbol{D}$ 进行 k 折交叉后的第 i 折数据,作为当前交叉验证的测试集;h^i 为利用剩余的 $k-1$ 折数据作为训练集得到的模型.

与 $\mathrm{RRMSE}(h,\boldsymbol{D}_{\mathrm{test}})$ 相同,$\mathrm{aRRMSE}(h,\boldsymbol{D})$ 的值越小,代表模型 h 的性能越好.

为了和其他算法进行公平的比较,与其他算法的实验设置相同,数据集 *rf*1 和 *rf*2 进行 5 折交叉验证,scm1d 和 scm20d 进行 2 折交叉验证,其余数据集利用 10 折交叉验证[1,7,8,10,11,13].

4.3 参数设置

本文的 TSF-TC 算法的参数设置如下:

TSF-TC 算法包含两个超参数:

(1)比率参数 $r\in(0,1]$,用于控制目标特定特征的数目. 根据式(3),r 越大,则每个目标的每个箱聚类数 K_i 就越大,则得到的中心点就越多,目标特定特征的维度就越大. 参照 LIFT[14] 算法,TSF-TC 的 r 设置为 0.1.

(2)阈值 $t\in[0,1)$,用于衡量考虑目标相关性的程度. 根据式(7),t 越大,则第二层模型堆叠的目标预测值就越少,与当前目标相关性越大. 为了避免相关性小的目标预测值降低模型的性能,阈值 t 设置为 0.9.

此外,XGBoost[18] 作为 TSF-TC 单一目标的学习器,针对不同规模的数据集,主要设置了以下参数:

(1)树的棵数:{100,300,500,1000}

(2)学习率:{0.01,0.03,0.06}

(3)树的最大层数:{3,5,10,15}

而对比算法 OKL、MROTS、RLC、SST、ERC、SVRCC 和 MMR 的参数与其论文中设置的参数相同[1,7,8,10,11,13].

4.4 假设检验

Friedman 检验[20] 可在多组数据集上进行多个算法性能的比较. Friedman 检验可提出假设 H_0:所有算法的性能相同. 当求得的概率 p 小于显著性水平 α 时可拒绝该假设,说明算法间存在显著性差异.

原假设 H_0 被拒绝,则需要进行后续 Nemenyi 检验[20] 进一步分析算法两两间的相对性能差异. 首先算出临界差异[20](Critical difference,CD):

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \quad (11)$$

式中,k 为比较算法的数量;N 为数据集的数量;q_α 为 Friedman 检验的临界值表中$(k-1,(k-1)(N-1))$位置的值.

CD 值是指两种算法被认为显著不同时所需的平均排名的最小差值,若两个算法的平均排名差值大于 CD,则相应的置信度 $1-\alpha$ 拒绝“两个算法性能相同”的假设. 一般 $\alpha=0.05$. Nemenyi 检验可用图形进行展

示,按照算法的平均排名在数轴上进行标注,若两个算法平均排名的差异小于 CD 值,则进行连线.

4.5　验证 TSF-TC 和 TSF 算法的预测性能

表4 展示了 TSF-TC 算法、TSF 算法与对比算法 OKL、MROTS、RLC、SST、ERC、SVRCC 和 MMR 的 aRRMSE,其中最佳结果用粗体表示,AveRank 表示算法的平均排名.

表 4　TSF-TC、TSF 算法与 7 个对比算法在不同数据集上的 aRRMSE

数据集 \ 算法	OKL	MROTS	RLC	SST	ERC	SVRCC	MMR	TSF	TSF-TC
andro	0.553	0.635	0.570	0.579	0.567	0.446	0.527	0.361	**0.339**
atp1d	0.364	0.404	0.384	0.372	0.372	0.378	**0.332**	0.356	0.347
atp7d	0.475	0.549	0.461	0.507	0.512	0.534	0.443	0.411	**0.410**
edm	0.741	0.812	0.735	0.740	0.741	0.698	0.716	0.685	**0.662**
enb	0.138	0.257	0.120	0.121	0.114	0.090	0.111	**0.086**	**0.086**
jura	0.599	0.625	0.596	0.591	0.590	0.589	0.582	0.582	**0.576**
oes10	0.432	0.558	0.419	0.421	0.420	0.354	0.403	0.346	**0.345**
oes97	0.535	0.605	0.523	0.524	0.524	0.464	0.497	0.442	**0.441**
osales	0.718	0.800	0.741	0.726	0.713	0.781	0.709	0.693	**0.688**
rf1	0.112	0.154	0.121	0.094	0.091	0.091	**0.085**	0.086	**0.085**
rf2	0.118	0.198	0.130	0.097	0.095	0.095	**0.086**	0.091	0.091
scm1d	0.342	0.449	0.345	0.336	0.330	0.328	0.324	0.323	**0.315**
scm20d	0.443	0.456	0.443	0.413	0.394	0.398	0.386	0.385	**0.381**
scpf	0.820	0.901	0.835	0.831	0.830	0.801	0.812	0.788	**0.786**
sf1	1.059	1.155	1.163	1.068	1.089	0.932	0.958	0.921	**0.916**
sf2	1.004	1.201	1.228	1.055	1.088	1.030	0.984	0.979	**0.961**
slump	0.699	0.778	0.690	0.695	0.689	0.556	0.587	**0.550**	**0.550**
wq	0.891	0.913	0.902	0.909	0.906	0.905	**0.889**	0.905	0.904
AveRank	6.111	8.889	6.667	6.389	5.667	3.889	3.333	2.222	**1.278**

从表4 中可以看到,TSF-TC 的平均排名为 1.278,TSF 的平均排名为 2.222,优于对比算法. 对表4 中展示的预测性能使用 Friedman 检验,验证算法间在统计上的显著性差异,提出原假设 H_0:这 9 个多目标回归算法是等价的,没有显著性差异. 表5 展示了 Friedman 检验的统计信息,其中数据集个数 $N=18$,对比算法个数 $k=9$,显著性水平为 $\alpha=0.05$.

表 5　Friedman 检验结果表

Friedman 检验	
Friedman 统计量	110.910
p-value($\alpha=0.05$)	2.475e−20
接受或拒绝	拒绝

参照表5 的结果,由于概率 p 远小于显著性水平,因此拒绝原假设 H_0,在统计上说明了算法间在预测性能上存在显著性差异. 原假设 H_0 被拒绝,则进行后续 Nemenyi 检验进一步分析算法两两间的相对性能差异. 由 $k=9$、$N=18$,查 Friedman 检验在 $\alpha=0.05$ 时的临界值表得到 $q_\alpha=2.007$,根据式(11)得到 CD = 1.832,因此若两两算法的平均排名差值大于 1.832 则表明这两个算法存在显著性差异.

图3 展示了 CD 值以及算法的平均排名,由图中可以看出,TSF-TC 性能最优,与排名第二的 TSF 的差异不大,但与排名第三的 MMR 有显著性差异;而 TSF 与排名第三的 MMR 差异不大. 由此说明,只考虑目标特征是不够的(TSF),只有结合目标特定特征和目标相关性

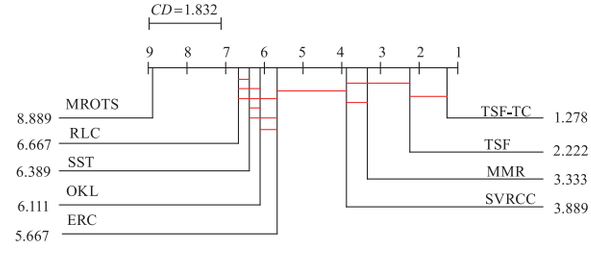

图3　TSF-TC、TSF与对比算法的平均排名

(TSF-TC)才能有效提高算法性能.

5 总结

本文提出了一种结合目标特定特征和目标相关性的多目标回归方法 TSF-TC. TSF 基于 K-Means 聚类分箱,再对分箱后的样本进行聚类分析构建目标特定特征;TSF-TC 在 TSF 的基础上,通过两层训练预测,堆叠与当前目标相关性大的目标预测值从而加入对目标相关性的考虑. 实验证明算法有效地提高了预测性能. TSF-TC 是分步考虑目标特定特征和目标相关性,在未来,将研究如何同时考虑目标特定特征和目标间相关性,从而进一步提升多目标回归的预测性能.

参考文献

[1] 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.

[2] Tsoumakas G, Katakis I. Multi-label classification: an overview[J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1 – 13.

[3] Zhang M L, Zhou Z H. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819 – 1837.

[4] Breiman L, Friedman J H. Predicting multivariate responses in multiple linear regression[J]. Journal of The Royal Statistical Society, 1997, 59(1): 3 – 54.

[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] Spyromitros-Xioufis E, Tsoumakas G, Groves W, et al. Multi-label classification methods for multi-target regression[J]. Computer Science, 2012, 1211(6581): 1 – 5.

[7] Tsoumakas G, Spyromitros-Xioufis E, Vrekou A, et al. Multi-target regression via random linear target combinations[A]. In Proceedings of 2014 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)[C]. Nancy: Springer, 2014. 225 – 240.

[8] Melki G, Cano A, Kecman V, et al. Multi-target support vector regression via correlation regressor chains[J]. Information Sciences, 2017, 415(1): 53 – 69.

[9] Wang J, Chen Z, Sun K, et al. Multi-target regression via target specific features[J]. Knowledge-Based Systems, 2019, 170(1): 70 – 78.

[10] Piyush R, Abhishek K, Daume H. Simultaneously leveraging output and task structures for multiple-output regression[A]. In Proceedings of Advances in Neural Information Processing Systems[C]. Lake Tahoe: MIT Press, 2012. 3194 – 3202.

[11] 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.

[12] Zhen X, Yu M, Zheng F, et al. Multitarget sparse latent regression[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1575 – 1586.

[13] Zhen X, Yu M, He X, et al. Multi-target regression via robust low-rank learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(2): 497 – 504.

[14] Zhang M. LIFT: Multi-label learning with label-specific features[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2015, 37(1): 107 – 120.

[15] Wong J A H A. Algorithm AS 136: A K-means clustering algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1): 100 – 108.

[16] Rendle S. Factorization machines[A]. In Proceedings of 2010 IEEE International Conference on Data Mining[C]. Sydney: IEEE Press, 2010. 995 – 1000.

[17] Michael P L V. Logistic regression[J]. Circulation, 2008, 117(18): 2395 – 2399.

[18] Chen T, Guestrin C. xgboost: A scalable tree boosting system[J]. Association for Computing Machinery, 2016, 10(1): 785 – 794.

[19] Coefficient P C. Pearson's correlation coefficient[J]. New Zealand Medical Journal, 1996, 109(1015): 38 – 39.

[20] 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

王 进 男,1979 年出生,工学博士,教授,主要研究方向为数据挖掘、机器学习.
E-mail: wangjin@cqupt.edu.cn

高选人 男,1994 年出生,硕士研究生,主要研究方向为机器学习与数据挖掘.
E-mail: 877906956@qq.com