基于小样本 SVR 的迁移学习及其应用

易 未[1],郑沫利[2],赵艳轲[2],毛 力[1],孙 俊[1]
(1. 江南大学 物联网工程学院,江苏 无锡 214122;
2. 国贸工程设计院,北京 100037)

摘 要:当前机器学习的技术已经运用到很多工程项目中,但大部分机器学习的算法只有在样本数量充足且运用在单一场景中的时候,才能获得良好的结果。其中,经典的支持向量回归机是一种具有良好泛化能力的回归算法。但若当前场景的样本数量较少时,则得到的回归模型泛化能力较差。针对此问题,以加权 ε 支持向量回归机为基础,提出了一种小样本数据的迁移学习支持向量回归机算法。该算法以加权 ε 支持向量回归机为 Bagging 算法的基学习器,使用与目标任务相关联的源域数据,通过自助采样生成多个子回归模型,采用简单平均法合成一个总回归模型。在 UCI 数据集和现实数据集——玉米棒与花生粒储藏环节损失数据集上的实验结果表明,该算法较标准 ε -SVR 算法与改进的 RMTL 算法在小数据样本上有更好的泛化能力。

关键词:支持向量回归机;迁移学习;加权 ε 支持向量回归机;Bagging;小样本数据
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2020)02-0047-05
doi:10.3969/j.issn.1673-629X.2020.02.010

Transfer Learning Based on Support Vector Regression Model for Small Sample Data and Its Applications

YI Wei[1],ZHENG Mo-li[2],ZHAO Yan-ke[2],MAO Li[1],SUN Jun[1]
(1. School of Internet of Things,Jiangnan University,Wuxi 214122,China;
2. Guomao Engineering Design Institute,Beijing 100037,China)

Abstract:Machine learning technologies have been applied to many industry programs nowadays,but most of them can obtain satisfied results with sufficient samples in a single situation. For instance,the classical support vector regression is a regression algorithm with better generalization ability. However,if the sample size in the current scene is small,the generalization ability of the regression model is poor. To solve this problem,we propose a transfer learning support vector regression algorithm for small sample data based on weighted ε support vector regression. In this paper, ε weighted support vector regression is taken as the basic learner of Bagging algorithm,and multiple sub regression models are generated by bootstrap using source data associated with target data,and a general regression model is synthesized by simple average method. Experimental results on the UCI datasets and the real dataset,the corn and peanut sales loss dataset,show that the proposed algorithm has better generalization ability than SVR algorithm and the improved RMTL algorithm on small data samples.

Key words:support vector regression;transfer learning;weighted ε support vector regression;Bagging;small sample

0 引 言

数据挖掘与机器学习技术在许多知识工程(例如分类、回归和聚类)等领域取得了有意义的成就[1]。但是,很多机器学习算法只有在训练集与测试集数据来源于单一场景、具有相同的特征空间和数据分布以及样本数量充足时,才能取得让人满意的实验结果。特别地,当样本数量不足时,容易出现过拟合现象,会显著地降低学习算法的效果[2]。然而在现实生活中,得到目标场景中的大量样本数据是很困难的,例如待观测的目标本身数量较少且不支持多次观测;或者是观测一次成本过高,只能使用现有的少量数据。

加权 ε 支持向量回归(ε -WSVR)算法具有扎实的理论基础和较好的泛化能力,应用在众多领域[3-5]。该算法为不同的样本设置不同的权值,给予样本不同

收稿日期:2019-03-07 **修回日期**:2019-07-10 **网络出版时间**:2019-11-07
基金项目:国家公益性行业科研专项(201513004);课题五(201513004-6)
作者简介:易 未(1993-),男,硕士研究生,研究方向为计算机应用与技术;郑沫利,教授级高级工程师,研究方向为粮食经济学。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0910.036.html

的精度要求和偏离精度要求的惩罚参数,减轻标准支持向量机对孤立点或噪声数据的敏感性,抑制过拟合现象的产生。

Bagging 算法通过对目标数据集进行数次有放回地抽样,形成多个不同的子数据集。然后在各个子数据集上使用基学习器,再将这些基学习器进行结合。对于分类任务采用简单投票法,对回归任务采用简单平均法,产生一个具有较强泛化能力的模型[6]。

迁移学习技术把目标任务称为目标域(target),与目标任务相关联的其他任务称为源域(domain),通过使用源域的数据或者知识来辅助建立目标域模型,提高模型的泛化能力[7]。针对分类任务,目前已经提出了很多基于迁移学习的研究成果,例如 Dai 等人借用 AdaBoosting 算法的思想提出了 TrAdaBoosting 算法,将源域中适合目标域模型训练的样本权重增加、其他样本权重减少[8];Liu 等人借用 Bagging 算法的思想提出了 BETL 算法,在自助采样后的子数据集上训练的初始分类器对未标示数据进行标示,然后用扩充了的标示数据训练未标示数据[9];Lin 等人提出 Double-bootstrapping 算法,在训练集上使用自助采样之后又再测试集上使用自助采样,两次自助采样提升模型分类精度[10]。针对回归任务,史荧中等人使用支持向量回归机(SVR)在历史数据和目标数据上构建两个尽可能相似的回归超平面,但该方法一旦运用到源域数据不为目标域的历史数据场景时,会给生成模型带来一定干扰[11];Yu 等人提出了改进的 RMTL 算法,使用 SVR 在源域与目标域数据上建立两个回归超平面模型,组合使用这两个模型提升泛化能力[12]。

但以上研究并没有对小样本(样本数小于 30)情况下的迁移回归进行研究。文中针对小样本数据情况下的回归系统建模问题,提出了一种小样本数据的迁移学习支持向量回归机方法,以加权 ε 支持向量回归机为 Bagging 算法的基学习器,使用与目标任务相关联的源域数据来弥补当前场景数据不足的问题。实验证明,该方法减少了回归误差,提高了目标模型的泛化能力。

1 相关知识

1.1 加权 ε 支持向量回归(ε-WSVR)算法

对于 ε 支持向量回归(ε-SVR)算法来说,设给定的训练样本集合为 $D = \{(x_1,y_1),\cdots,(x_i,y_j),\cdots,(x_m,y_m)\}$,$x_i \in R^N$,$y_i \in R$,$\varepsilon$-SVR 算法的基本思想是得到一个形如式(1)的回归模型:

$$f(x) = \omega^{\mathrm{T}} \cdot \Phi(x) + b \tag{1}$$

其中,ω 与 b 是模型的参数,Φ 是一个非线性映射,将有限维 x 映射到一个高维特征空间使训练样本线性可分,$\Phi(x)$ 为将 x 映射后的特征向量。可以使用适当的核函数 $\kappa(x_i,x_j)$,使 x_i 与 x_j 在高维特征空间的内积等于其在原始样本空间上内积的结果。

传统回归模型通常当且仅当模型的输出 $f(x)$ 与真实值 y 相等时,损失才为零,但是 ε-SVR 仅当 $|f(x) - y| < \varepsilon$ 时,损失才为零。于是 ε-SVR 可形式化为式(2):

$$\min_{\omega,b,\xi_i,\xi_1} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_1) \tag{2}$$

约束条件为:

$$f(x_i) - y_i \leqslant \varepsilon + \xi_i$$

$$y_i - f(x_i) \leqslant \varepsilon + \xi_i$$

$$\xi_i \geqslant 0, \xi_i \geqslant 0, i = 1,2,\cdots,m \tag{3}$$

式(2)中,第一项为找到最小的 ω 使 $\frac{2}{\|\omega\|}$ 最大。第二项 $C > 0$ 为常数,当 C 为无穷大时,使所有训练样本满足式(3);当 C 为常数时,允许一部分训练样本不满足式(3)。

加权 ε 支持向量回归(ε-WSVR)算法[13]在式(2)的基础上添加权值 μ,得到式(4):

$$\min_{\omega,b,\xi_i,\xi_1} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\mu_i(\xi_i + \xi_1) \tag{4}$$

式(4)可以使用 SMO(sequential minimal optimization)算法求解,算法的时间复杂度为 $O(n^{2.3})$[14]。

1.2 ε-WSVR 算法权值的确定

源域中的样本数据与目标域中的样本数据越相似,则辅助建立的目标域模型泛化能力越好[8]。文中使用样本间的标准化欧氏距离来定义样本数据的相似情况,则 ε-WSVR 算法权值 μ 按照式(5)来计算:

$$\mu_j = \frac{1}{\min(l) + 1} \tag{5}$$

其中,l 为源域中样本 x_i 到目标域中样本标准化欧氏距离的列向量,$\min(l)$ 为该列向量中的距离最小值。所以权值 μ 的取值范围是(0,1]。

1.3 改进的 RMTL 算法

设源域样本集合为 s,目标域样本集合为 t,改进的 RMTL 问题可形式化为:

$$\min_{\omega_s,\omega_t,b_s,b_t} \frac{1}{2}\|\omega_s\|^2 + \frac{1}{2}\|\omega_t\|^2 + \frac{1}{2}\|\omega_s - \omega_t\|^2 + C_s\sum_{i=1}^{m_s}(\xi_{s,i} + \xi_{s,1}) + C_t\sum_{i=1}^{m_t}(\xi_{t,i} + \xi_{t,1}) \tag{6}$$

约束条件为:

$$f_s(x_{s,i}) - y_{s,i} \leqslant \varepsilon_s + \xi_{s,i}$$

$$y_{s,i} - f_s(x_{s,i}) \leqslant \varepsilon_s + \xi_{s,i}$$

$$f_t(x_{t,i}) - y_{t,i} \leqslant \varepsilon_t + \xi_{t,i}$$

$$y_{t,i} - f_t(x_{t,i}) \leqslant \varepsilon_t + \xi_{t,i}$$

$$\xi_{s,i}, \xi_{s,i}, \xi_{t,i}, \xi_{t,i} \geqslant 0 \tag{7}$$

其中 $\lambda > 0$ 为常数,当 λ 较大时,将会导致源域与目标域的回归向量 $\boldsymbol{\omega}$ 相似;当 λ 较小时,将会导致源域与目标域的回归向量 $\boldsymbol{\omega}$ 不同。

求解式(6)与式(7)的对偶问题,可得该算法模型为:

$$y_t = \frac{\lambda}{1 + 2\lambda} \sum_{i=1}^{m_s} \boldsymbol{\omega}_s^{\mathrm{T}} x + \frac{1 + \lambda}{1 + 2\lambda} \sum_{i=1}^{m_t} \boldsymbol{\omega}_t^{\mathrm{T}} x + b_t \tag{8}$$

2　小样本数据的迁移学习支持向量回归算法

2.1　算法的基本思想

小样本数据的迁移学习支持向量回归算法的主要思想是将 ε -WSVR 算法作为 Bagging 算法的基学习器,使用自助采样(bootstrap)方法从源域和目标域数据集中进行采样,得到一系列子数据集。然后计算子数据集各个样本到目标域数据集中的标准化欧氏距离,得到子数据集到目标域数据集的最小距离,并把这个距离加一的倒数作为 ε -WSVR 算法的权值。最后使用这一系列子学习器对测试数据进行计算,把子学习器结果的简单平均值作为小样本数据的迁移学习支持向量回归算法的结果。

2.2　算法流程

输入:源域数据集 source,目标域数据集 target,Bagging 算法的基学习器个数 T,ε -WSVR 算法 ζ。

过程:

1:for $t = 1, 2, \cdots, T$ do

2: C = standardizedEuclideanDist(D_{bs}, target)

D_{bs} 是 source 与 target 上自助采样产生的样本集合

3: $C' = \min(C)$

C' 是 D_{bs} 到 target 数据集距离的最小值

4: $h_t = \zeta(D_{\mathrm{bs}}, \mathrm{target}, C')$

ε -WSVR 算法 ζ 的权值为 $\mu = \dfrac{1}{\min(C') + 1}$

5:end for h_i

输出: $H(x) = \dfrac{1}{T} \sum_{i=1}^{T} h_i(x)$ 。

2.3　算法的时间复杂度分析

假设 source 数据集大小为 n,target 数据集大小为 m。Bagging 算法的基学习器 ε -WSVR 算法的复杂度为 $O((m+n)^{2.3})$,自助采样过程的复杂度为 $O(m+n)$,计算到 target 数据集距离复杂度为 $O(mn)$,计算权值复杂度为 $O(m+n)$。小样本数据的迁移学习支持向量回归算法的复杂度为 $T*(2*O(m+n) + O(mn) + O((m+n)^{2.3}))$,考虑到 T 通常是一个不太大的常数,因此,小样本数据的迁移学习支持向量回归算法的时间复杂度为 $O((m+n)^{2.3})$,与直接使用 ε -WSVR 算法的复杂度同阶。

3　实　验

文中实验将使用辽宁、陕西、山西、安徽、江苏、湖北、湖南、四川和广东一共九个省份的大米与玉米储藏环节损失情况调查数据,以及三个 UCI Machine Learning Repository 数据集(分别是 Wine Quality、Student Performance、PM2.5 Data of Five Chinese Cities)对提出的小样本数据的迁移学习支持向量回归算法进行实验,其中将四个目标域数据集按照 2∶1 的比例划分为训练集与测试集。将分别构建以下回归模型进行对比:(1)只使用目标域数据和标准 ε -SVR 算法建立的回归模型(SVR-t);(2)利用源域数据和目标域数据与标准 ε -SVR 算法建立的回归模型(SVR-s,t);(3)使用改进的 RMTL 算法基于源域数据和目标域数据建立的回归模型(RMTL-s,t);(4)使用文中方法基于源域数据和目标域数据建立的回归模型(bagg-WSVR)。以上四种回归模型采用式(9)均方误差[15]进行比较:

$$\mathrm{MSE} = \frac{1}{k} \sum_{i=1}^{k} (y_i - y_1)^2 \tag{9}$$

其中,y_i 为实际值,y_i 为预测值,k 为测试样本数量。

SVR-t、SVR-s,t、bagg-WSVR 这三组模型核函数均选择高斯核函数,参数 C 与 $g(g = \dfrac{1}{2 * \sigma^2})$ 使用网格搜索确定,C 的取值范围为 2^-8 到 2^8;g 的取值范围为 2^-8 到 2^8。RMTL-s,t 模型参数 C 与 λ 使用网格搜索确定,C 的取值范围为 2^-8 到 2^10;λ 的取值范围为 0.1 到 10。bagg-WSVR 基学习器个数为 10。

实验环境:实验硬件为 Intel Core i5-2430M CPU,主频 2.40 GHz,内存 8 GB,编程环境为 Matlab R2016b 与 MyEclipse2015。

3.1　四个数据集描述

Wine Quality 数据集使用白葡萄酒当作源域数据集,从红葡萄酒数据集中随机选择 30 条样本数据作为目标域数据集,quality 属性作为输出变量;Student Performance 数据集使用 GP 学校的 Math 当作源域数据集,从 MS 学校的 Math 数据集中随机选择 30 条样本数据作为目标域数据集,G_1、G_2 与 G_3 属性的数值之和作为输出变量;PM2.5 Data of Five Chinese Cities 数据

集使用北京 2014 年 8 月删除时间属性当作源域数据集,从沈阳 2014 年 8 月删除时间属性数据集中随机选择 30 条样本数据作为目标域数据集,PM_USPost 属性作为输出变量;粮食储藏数据集使用大米数据当作源域数据集,玉米数据作为目标域数据集,储藏环节损失率作为输出变量。四个数据集描述如表 1 所示。

表 1 四个数据集描述

数据集	源域数据数	目标域数据数
wine	4 898	30
student	349	30
pm2.5	733	30
粮食储藏	125	26

3.2 四个数据集上各个模型的回归结果与分析

wine 数据集上各个模型的性能比较如表 2 所示。

student 数据集上各个模型的性能比较如表 3 所示,

pm2.5 数据集上各个模型的性能比较如表 4 所示。

粮食储藏数据集上各个模型的性能比较如表 5 所示。

表 2 wine 数据集上各个模型的性能比较

实验方法	训练样本数	MSE	参数选择
SVR-t	20	3.057E-4	$C=2^3, g=2^{-4}$
SVR-s,t	4 918	2.696E-4	$C=2^2, g=2^{-8}$
RMTL-s,t	4 918	3.108E-3	$C=2^{10}, \lambda=1$
bagg-WSVR	4 918	2.308E-4	$C=2^1, g=2^{-6}$

表 3 student 数据集上各个模型的性能比较

实验方法	训练样本数	MSE	参数选择
SVR-t	20	7.001E-3	$C=2^8, g=2^2$
SVR-s,t	369	5.642E-3	$C=2^{-1}, g=2^2$
RMTL-s,t	369	1.891E-2	$C=2^{10}, \lambda=0.1$
bagg-WSVR	369	3.989E-3	$C=2^6, g=2^1$

表 4 pm2.5 数据集上各个模型的性能比较

实验方法	训练样本数	MSE	参数选择
SVR-t	20	2.375E-3	$C=2^8, g=2^8$
SVR-s,t	753	2.010E-3	$C=2^8, g=2^1$
RMTL-s,t	753	3.163E-3	$C=2^{10}, \lambda=1$
bagg-WSVR	753	1.802E-3	$C=2^6, g=2^8$

表 5 粮食储藏数据集上各个模型的性能比较

实验方法	训练样本数	MSE	参数选择
SVR-t	20	8.164E-3	$C=2^8, g=2^{-4}$
SVR-s,t	145	1.104E-2	$C=2^8, g=2^{-1}$
RMTL-s,t	145	7.443E-3	$C=2^{10}, \lambda=1$
bagg-WSVR	145	7.383E-3	$C=2^5, g=2^{-2}$

由表2 到表5 可以看出:(1)当源域与目标域数据关联程度很大时,利用源域数据和目标域数据与标准 ε -SVR 算法建立的回归模型较只使用目标域数据和标准 ε -SVR 算法泛化性能好;当源域与目标域数据有关联,但是关联程度不太大时,利用源域数据和目标域数据与标准 ε -SVR 算法建立的回归模型较只使用目标域数据和标准 ε -SVR 算法泛化性能差,出现了“负迁移”现象。(2)改进的 RMTL 算法在小样本数据情况下的算法性能很不稳定,原因是目标域样本较少,导致标准 ε -SVR 算法建立的回归模型过拟合。(3)文中提出的算法在四个数据集上有着更好的泛化性能,因为文中算法根据源域中与目标域样本的相似程

度,给相似样本赋予更大的权重数值;同时训练样本数目较大,防止生成模型过拟合现象产生,从而提高了泛化性能。

4　结束语

针对小样本数据情况下的回归系统建模问题,提出了一种小样本数据的迁移学习支持向量回归机建模方法。以加权 ε 支持向量回归机为 Bagging 算法的基学习器,使用与目标任务相关联的源域数据,通过自助采样生成多个子回归模型,采用简单平均法合成一个总回归模型。通过 UCI 数据集与现实数据集——玉米棒与花生粒储藏环节损失数据集的验证,结果表明该算法较标准 ε -SVR 算法与改进的 RMTL 算法在小数据样本上有更好的表现。但该算法也有一些不足之处:由于采用 Bagging 算法思想,有放回的抽样产生子数据集造成算法回归结果不稳定。下一阶段将改进子数据集产生的抽取方法,使源域中与目标域相似的样本更容易被选取,降低结果的不稳定性。

参考文献:

[1] PAN S, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345-1359.

[2] 杨道军,王　冉,沈　刚. SVM 与 ANN 在湖泊富营养化评价中的对比研究[J]. 环境科学与技术,2012,35(1):173-177.

[3] HAN X, CLEMMENSEN L. On weighted support vector regression[J]. Quality & Reliability Engineering International, 2014,30(6):891-903.

[4] ZHANG Y, XU S, CHEN K, et al. Fuzzy density weight-based support vector regression for image denoising[J]. Information Sciences, 2016, 339:175-188.

[5] DAI Z, WANG L, CHEN Y, et al. A pipeline for improved QSAR analysis of peptides: physiochemical property parameter selection via BMSF, near-neighbor sample selection via semivariogram, and weighted SVR regression and prediction [J]. Amino Acids, 2014, 46(4):1105-1119.

[6] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[7] 庄福振,罗　平,何　清,等. 迁移学习研究进展[J]. 软件学报,2015,26(1):26-39.

[8] DAI W, YANG Q, XUE G, et al. Boosting for transfer learning [C]//Proceedings of the 24th international conference on machine learning. Corvalis, Oregon, USA: [s. n.], 2007:193-200.

[9] LIU X, WANG G, CAI Z, et al. Bagging based ensemble transfer learning[J]. Journal of Ambient Intelligence & Humanized Computing, 2016, 7(1):29-36.

[10] LIN D, AN X, ZHANG J. Double-bootstrapping source data selection for instance-based transfer learning [J]. Pattern Recognition Letters, 2013, 34(11):1279-1285.

[11] 史荧中,王士同,蒋亦樟,等. 迁移学习支持向量回归机[J]. 计算机应用,2013,33(11):3084-3089.

[12] YU B, JI H. Near-infrared calibration transfer via support vector machine and transfer learning[J]. Analytical Methods, 2015, 7(6):2714-2725.

[13] 孙德山,吴今培,侯振挺,等. 加权支持向量回归算法[J]. 计算机科学,2003,30(11):38-39.

[14] SHEVADE S K, KEERTHI S S, BHATTACHARYYA C, et al. Improvements to the SMO algorithm for SVM regression[J]. IEEE Transactions on Neural Networks, 2000, 11(5):1188-1193.

[15] 王志明,谭显胜,袁哲明,等. 自调用支持向量回归和偏最小二乘优化支持向量机参数[J]. 小型微型计算机系统,2010,31(9):1815-1819.