ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 414－421
DOI:10. 16337/j. 1004-9037. 2019. 03. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于迁移学习的电力通信网异常站点业务数量预测

杨济海[1]　李号号[2]　彭汐单[3]　张智成[4]　黄　倩[2]　李石君[2]

(1. 国网江西省电力有限公司信息通信分公司,南昌, 330077;2. 武汉大学计算机学院,武汉, 430072;3. 国网江西省电力有限公司,南昌, 330077;4. 南瑞集团有限公司,南京, 210003)

摘　要: 现有的多源迁移学习算法对回归问题的研究极少,大多是解决对称的二分类问题,本文提出了加权多源TrAdaBoost的回归算法,其中误差容忍系数能一定程度解决源领域样本权重缩减过快的问题,提高了算法的效果。在修改后的Friedman #1回归问题上进行了实验,验证了该算法的有效性,误差容忍系数可以提高大约0.01的R^2分数。将该算法应用到电力通信网的行业问题中,提出了异常站点(业务数量缺失严重的站点)检测与真值预测模型,在特征工程中使用了社交网络分析的方法,充分考虑了站点在拓扑图中的重要性。最终的实验效果进一步验证了算法的有效性。

关键词: 机器学习;电力通信网;回归算法;多源迁移学习;异常检测

中图分类号: TP18;TN915　　**文献标志码:** A

Method Based on Transfer Learning for Predicting Quantity of Service in Power Communication Network

Yang Jihai[1], Li Haohao[2], Peng Xidan[3], Zhang Zhicheng[4], Huang Qian[2], Li Shijun[2]

(1. Information & Telecommunication Branch, State Grid Jiangxi Electric Power Company, Nanchang, 330077, China; 2. School of Computer Science, Wuhan University, Wuhan, 430072, China; 3. State Grid Jiangxi Electric Power Company, Nanchang, 330077, China; 4. NARI Group Corporation, Nanjing, 210003, China)

Abstract: The existing multi-source transfer learning algorithms have very few researches on regression problems, and most of them are symmetric two-class classification problems. This paper presents a weighted multi-source TrAdaBoost regression algorithm, in which the error tolerance coefficient is proposed to solve the problem that the sample weight of the source domain is reduced too quickly, thus the effect of the algorithm is improved. Experiments are performed on the modified Friedman #1 regression problem to verify the effectiveness of the algorithm. The error tolerance coefficient can increase the R^2 score by approximately 0.01. In this paper, the proposed algorithm is applied to the industry problems of power communication networks, and the anomaly site (sites with a large number of missing services) detection and true value prediction models are proposed. Moreover, the social network analysis methods are used in the feature engineering, and the importance of the site in the topology is fully considered. Finally, experimental results verify the effectiveness of the algorithm.

Key words: machine learning; power communication networks; regression algorithm; multi-source transfer learning; anomaly detection

收稿日期:2018-07-23;**修订日期:**2019-04-19

引　言

迁移学习是机器学习的一个新方向,它是利用已存在的知识对不同但相关的领域进行训练学习。迁移学习打破了传统机器学习所需满足的条件——训练数据和测试数据独立同分布,以及拥有足够的数据来训练一个好的模型[1]。研究表明,两个不同领域的相似度越高,迁移学习就越容易,效果越好,否则往往效果不佳,甚至出现"负迁移"的结果。迁移学习已经成功应用到多个领域[2],如文本情感分析、图像分类、人类活动识别、软件缺陷分类和多语言文本分类等。表1总结了现有的各种迁移学习方法。

表1　各种迁移学习方法的举例和描述

Tab. 1　Examples and descriptions of various migration learning methods

	方法	描述
基于关系的迁移	Second-order Markov Logic[3]	利用源域学习逻辑关系网络应用到目标域
基于实例的迁移	TrAdaBoost[4], Kernel Mean Matching, Density ratio estimation[5]	对源域的实例进行加权
基于特征的迁移	TCA[6], SFA[7], GFK, TKL	将两个域变换到同一特征空间
基于模型的迁移	TransEMDT[8], TRCNN[9], TaskTrAdaBoost[10]	将源域的模型用到目标域

Dai等提出了基于实例的TrAdaBoost[4]算法,该算法的思想是最大限度利用源数据,找到源数据中与目标数据相关的数据,然后和目标数据一起训练学习。但是TrAdaBoost算法只利用了单个源数据,算法的结果依赖于源数据与目标数据的相关性,如果相关性很弱,容易产生负迁移。Cheng等人通过考虑多个源与目标的相关性,提出了两种多源学习算法,MTrA和TTrA[11]。多源的迁移学习主要研究当源领域为多个时如何进行迁移的问题,主要的成果有Transitive transfer learning[12](两个相似度不高的域利用从第三方中学习到的相似度关系,完成知识的传递迁移),Distant domain TL[13](在相似度极低的两个域进行迁移时,用Autoencoder自动从多个中间辅助域中选择知识)等,多源的迁移可以有效地利用多个领域的知识,综合起来达到较好的效果。在多源的迁移学习问题中,现有的算法如MultiSource-TrAdaBoost, Task-TrAdaBoost, Weighted multi-source TrAdaBoost等研究的都是对称的二分类问题,缺少对回归问题的研究,回归问题和二分类问题在目标函数的定义等方面还是有着很大的区别,所以本文将提出加权多源TrAdaBoost的回归算法。

本文除了关注迁移学习的理论研究外,还将重点关注其在电力通信网中的应用。国家电网通信管理系统(TMS)中普遍存在账物与实物不一致、数据录入错误和数据缺失的问题。阮筠萃介绍了电力通信网管理系统的静态资源与实际不符合、动态资源关联错误、基础数据保险不到位的各类型数据质量问题,以及这些问题带来的巨大挑战[14]。Liu等详细分析了导致电力数据质量问题的原因[15]。在TMS系统中存在着一些业务记录严重缺失的站点,本文将这些站点定义为异常站点。使用本文提出的加权多源TrAdaBoost的回归算法对站点正确的业务数量进行预测,将极大地降低数据维护的成本,具有巨大的现实意义。

1　加权多源TrAdaBoost的回归算法

在分类的问题中,对于一个样本x_i的预测h_t,要么是正确的,要么是不正确的,所以其误差$e_i=$

$|y_i - h_t(x_i)|$为0或者1。在回归问题中,该误差有可能极其大,所以需要将其归一化,这样才能使用TrAdaBoost的权重更新机制。本文将AdaBoost的误差函数引入进来,误差函数有如下3种方式,其中D表示最大的误差。

$$e_i = |y_i - h_t(x_i)|$$

$$D = \max_{i=0}^{n} |e_i|$$

$$e'_i = \begin{cases} e_i/D & \text{线性} \\ e_i^2/D^2 & \text{平方} \\ 1-\exp(-e_i/D) & \text{指数} \end{cases}$$

在TrAdaBoost的权重更新机制中,由于源领域样本的权值只会减少而不会增加,而目标领域样本的权值只会增加却不会减少,因此两个领域的权重之差会越来越大,在迭代次数较多的情况下,会严重影响模型的效果。该问题在回归问题中显得尤为严重,因为e_i几乎不可能为0,即使很小的误差,也会导致权重的缩减,所以在多次迭代之后,源领域的样本权重很有可能缩减为0。为了解决该问题,本文引入了误差容忍系数γ,如果某个样本的误差小于容忍系数,那么其权重不发生变化,如果该样本的误差超过了容忍系数,其权重才会发生变化。误差容忍系数将一定程度解决源领域权重缩减的问题,提高迁移的效果,并控制模型对误差的容忍程度。

在多源的加权学习中,每个源领域将得到一个弱的学习器,根据弱学习器在目标领域上的误差情况,可以每次只选取误差最小(相关性最强)的源领域,也可以根据该误差的大小对每个弱学习器进行加权,加权之后进行集成,得到本次迭代中的测试误差,根据该测试误差去调整样本的权重。前者的做法往往会造成其他的源领域样本失去辅助作用,所以本文选择后者的做法。部分符号说明见表2。

表2 符号表

Tab. 2 Symbol table

符号	详细描述
$(S_1, S_2, \cdots, S_N)$	N个源领域
$(x_1^{s_k}, x_2^{s_k}, \cdots, x_{n_{S_k}}^{s_k})$	第S_k个源领域的实例,n_{S_k}为实例数量
$(y_1^{s_k}, y_2^{s_k}, \cdots, y_{n_{S_k}}^{s_k})$	第S_k源领域的实例的标签
$(w_1^{s_k}, w_2^{s_k}, \cdots, w_{n_{S_k}}^{s_k})$	第S_k源领域实例的权重
$(x_1^{\text{target}}, x_2^{\text{target}}, \cdots, x_{n_{\text{target}}}^{\text{target}})$	目标数据源的实例
$(y_1^{\text{target}}, y_2^{\text{target}}, \cdots, y_{n_{\text{target}}}^{\text{target}})$	目标数据源的实例标签
T	算法的迭代次数
$(D_{S_1}, \cdots, D_{S_k}, \cdots, D_{S_N})$	N个源领域训练集
D_{target}	目标领域训练集
γ	容忍系数

算法1 加权多源TrAdaBoost的回归算法

输入 $(D_{S_1}, \cdots, D_{S_k}, \cdots, D_{S_N}), D_{\text{target}}, T, \gamma$

输出 回归器$f(\bullet): X \to Y$

步骤:

(1) 初始化参数$\phi(s)$,其中$n_s = \sum_{k=1}^{N} n_{s_k}$,$n_s$为所有源领域的样本数的总和

$$\phi(s) = \frac{1}{1+\sqrt{2\ln(n_s)/T}}$$

(2) 初始化样本权重。

(3) for t:1 → T。

(4) 合并训练数据集$D_k=(D_{S_k},D_{\text{target}})$,同时对权重向量$(w^{S_k},w_{\text{target}})$归一化。

(5) 对合并后的每一个训练数据集D_k,$(w^{S_k},w_{\text{target}})$,将其代入到回归模型。此处选择带参数的SVR模型(也可以选择回归树等),即

$$h_t^k(x^{S_k})=\sum_{i=1}^{n_{S_k}}(\alpha_i^{S_k}-\alpha_i^{'S_k})K(x_i^{S_k},x^{S_k})+b^{S_k}\quad k=1,2,\cdots,N$$

(6) 计算h_t^k在D_{target}上误差,此处采用的是线性的误差函数

$$\varepsilon_t^k=\sum_{i=1}^{n_{target}}w_i|y_i-h_t^k(x_i)|/D\quad D=\max_{i=1}^{n_{\text{target}}}|y_i-h_t^k(x_i)|$$

(7) 根据误差ε_t^k更新预测模型h_t^k的权重$\delta_t^k=\dfrac{e^{1-\varepsilon_t^k}}{e^{\varepsilon_t^k}}$,然后将多个弱回归器进行集成得到第$t$次迭代的候选回归模型

$$h_t=\sum_{k=1}^{N}\frac{\delta_t^k}{\sum_{k=1}^{N}\delta_t^k}h_t^k$$

(8) 计算在第t次迭代得到的候选预测模型h_t在D_{target}上的误差

$$\varepsilon_t=\sum_{i=1}^{n_{\text{target}}}w_i|y_i-h_t(x_i)|/D\quad D=\max_{i=1}^{n_{\text{target}}}|y_i-h_t(x_i)|$$

(9) 更新所有样本的权重。根据上面得到的误差ε_t更新各个源领域和目标领域样本的权重。对于目标领域中的样本,需要根据预测的误差大小增加样本的权重,即表示对于预测错误的样本,应增加此样本的重要性,达到强调此样本的目的。对于源领域中的样本,如果样本预测值小于γ,则此样本的权重不变;相反,如果相差大于γ,则减小此样本的权重,即表示预测错误的样本对目标数据的学习没有帮助,应该降低这些样本的影响。设置ϕ_t为

$$\phi_t=\frac{\varepsilon_t}{1-\varepsilon_t}\quad 0\leqslant\varepsilon_t\leqslant 0.5$$

更新目标领域样本的权重

$$e_i=|y_i-h_t(x_i)|/D\quad D=\max_{i=1}^{n_{\text{target}}}|y_i-h_t(x_i)|$$

$$w_{i(t+1)}=\begin{cases}w_{it} & \varepsilon_t=0\\ w_{it}\phi_t^{1-e_i} & 0<\varepsilon_t\leqslant 0.5\end{cases}\quad i=1,2,\cdots,n_{\text{target}}$$

更新各个源领域样本的权重

$$e_i^{s_k}=\left|y_i^{s_k}-h_t^k(x_i^{s_k})\right|/D\quad D=\max_{i=1}^{n_{s_k}}\left|y_i^{s_k}-h_t^k(x_i^{s_k})\right|$$

$$w_{i(t+1)}^{s_k}=\begin{cases}w_{it}^{s_k}\phi(s)^{e_i^{s_k}} & \left|y_i^{s_k}-h_t{}^k(x_i^{s_k})\right|>\gamma\\ w_{it}^{s_k} & \left|y_i^{s_k}-h_t{}^k(x_i^{s_k})\right|\leqslant\gamma\end{cases}\quad k=1,2,\cdots,N;\ i=1,2,\cdots,n_{s_k}$$

(10) 如果$t\leqslant T$,则转到步骤(4);如果$t>T$,则计算出最终的预测模型$f(x)$。$f(x)$为$h_t(x)$的加权中位数,用$\ln(1/\phi_t)$作为$h_t(x)$的权值。

2 通信网异常站点检测与真值预测

异常站点的挖掘与纠正是国家电网根据实际的需求提取出来的研究课题,具有很大的研究价值和

现实意义。在通信网管理系统中,小部分站点存在系统录入的业务数量与实际的业务数量相差较大的问题,本文将这种现象称为业务数量缺失,将业务数量缺失的站点称为异常站点。这类问题给实际的运营和管理带来巨大的困难,单靠人力无法完成这类数据质量的治理。

在对业务数量进行预测时,由于全国各省的站点数量不同——大的省份有两千多个站点,而小的只有两百个左右的站点,各省经济、人口、基础设施等因素的不同,导致了其数据分布也各不相同。如果放到同一个模型中进行训练,无法得到满意的预测精度。本文通过上述的加权多源TrAdaBoost的回归算法,运用其他省份的数据来训练目标数据,使模型能够达到较为满意的效果。本文设计的异常站点检测与真值预测的模型如图1所示。

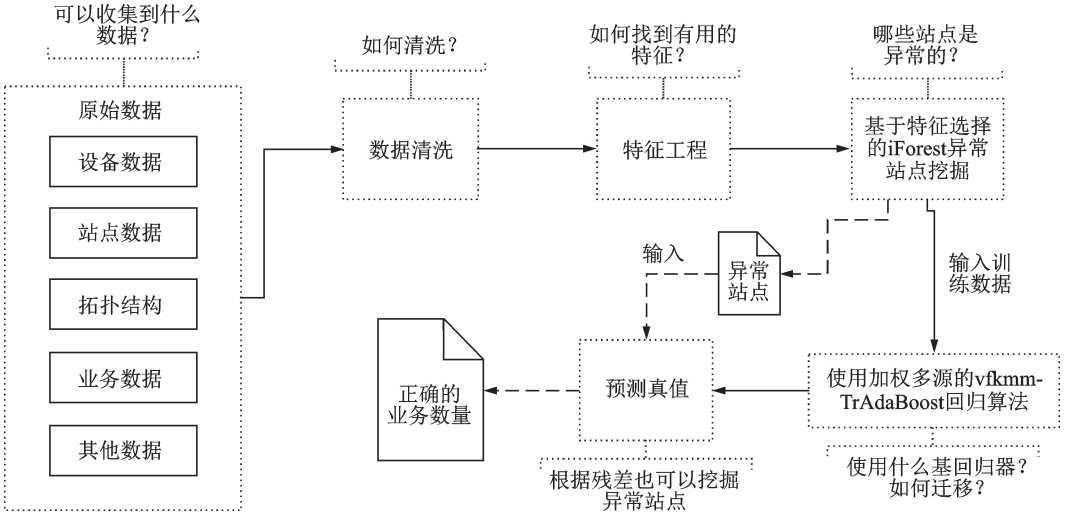

图1 异常站点检测与真值预测模型图

Fig.1 Abnormal site detection and true value prediction model

数据清洗是对初始数据进行缺失值处理、异常值处理和冗余去除等操作。在特征工程中,选取的特征主要来源于业务专家推荐和相关性分析方法,本文引入社交网络分析的方法,提取中心度和PageRank值的特征,充分考虑了站点在拓扑结构中的重要度。使用图G来表达某个省份的站点与光缆的拓扑图,站点之间的权重为光缆的数量,以邻接矩阵的方式来表达,其中g为站点的数量,如果i,j之间无光缆连接,则$x_{ij}=0$。$C_D(i)$为站点i的中心度。图2绘制了中心度与业务数量的曲线图,也能验证中心度与业务数量的强关联关系。

$$G=\begin{pmatrix} x_{11} & \cdots & x_{1g} \\ M & O & M \\ x_{g1} & L & x_{gg} \end{pmatrix}$$

$$C_D(i)=x_{ij} \quad i\neq j$$

本文为了引入其他站点对该站点业务数量的影响,使用了PageRank算法提取特征,节点x_i的PageRank值的更新公式如式(1)所示,其中$M(x_i)$为所有站点对站点x_i有出链的站点集合,$L(x_j)$表示x_j到所有其他站点出链的权重之和,x_{ji}表示站点x_j到x_i的权重,d为阻尼系数,一般经验值取0.85。式(1)表明,如果与站点x_i相连接的站点个数越多,则说明该站点的重要性越高,如果与该站点连接的其他站点越重要,那么该站点很可能也越重要,承载的业务越多,如果该站点所占连接源的站点的权重越高,则说明该站点重要性越高。除了上述特征外,还有站点类型、电压等级、端口占有率、调度等级、建成年

限、设备数量和机房数量等特征。

$$PR(x_i)=\frac{1-d}{g}+d\cdot\sum_{x_j\in M(x_i)}\frac{x_{ji}}{L(x_j)}PR(x_j) \tag{1}$$

异常站点的挖掘首先是一个异常检测问题,有很多从概念视角和具体应用视角的异常检测相关的综述[16-18]。本文选择了无监督算法iForest选择初始的异常站点,使用iForest主要是因为该算法需调节的参数少、准确率高、运行效率快。剔除异常站点后剩下的站点就可以被输入到回归模型中。如果不过滤这些离群点数据,将会严重影响模型最终的效果和泛化能力。如果要预测某个省份的站点业务数量,就将该省份的数据作为目标领域,其他省份的数据作为源领域。使用上述的加权多源TrAdaBoost的回归算法,解决了不同省份之间数据分布不同、部分省份训练数据过少的问题。使用带参数的SVR作为基回归器,因为SVR采用的是结构化风险最小化,更适合小样本的使用场景且有更好的鲁棒性。异常站点来自两个部分,一是iForest发现的异常站点(离群点),二是回归模型的预测值与观测值残差较大的站点。

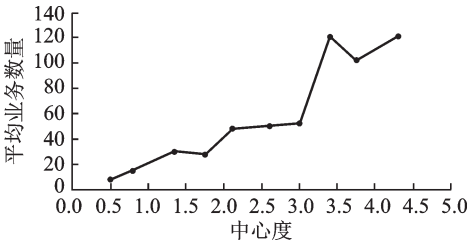

图2 站点中心度与业务平均数量的关系

Fig.2 Relationship between site center degree and average number of businesse

3 实 验

3.1 Friedman #1 回归问题

Friedman #1 (Friedman, 1991)是一个非常著名的回归问题,为了验证提出的加权多源TrAdaBoost回归算法,本文修改了Friedman问题,以便将其使用到迁移学习的问题中。通过式(2)可生成目标领域和源领域的数据集,其中N表示正态分布。为生成源领域的数据集,本文对式(2)的参数a_i,b_i,c_i进行了改进,不再取固定值,a_i,b_i由$N(0,0.1d)$生成,c_i由$N(0,0.05d)$生成,这里的d是一个参数,控制目标领域样本与源领域样本的相似度。本文使用不同的d随机生成了5个源领域的数据集,d的取值为[0.5,1]。每个源领域数据集的大小为200,目标训练集的大小为15,测试数据集的大小为600,生成了60份目标领域和源领域的数据集,分为3组。图3展示了生成的数据集,其中不同的数据集以不同的颜色表示。由于目标数据集的数据较少(黑色的点),难以训练出有效的回归模型。测试数据集的分布用蓝线表示,可以观察到源领域的数据分布(红色、黄色等)明显与目标领域不同,因此改造后的Friedman回归问题是非常适合迁移学习的使用场景。

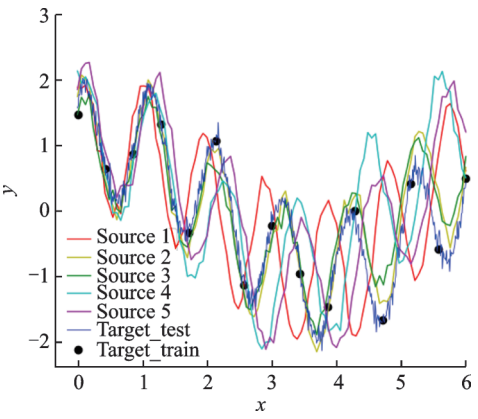

图3 各个领域数据分布图

Fig.3 Distribution of data in various fields

$$y=a_1*\sin(b_1*x+c_1)+a_2*\sin(b_2*6x+c_2)+N(0,0.1) \tag{2}$$

考虑以下几组对比实验:

(1) AdaBoostRegressor表示将所有的源领域数据和目标领域的数据进行合并,不使用迁移学习;

(2) Muti-TrAdaBoost表示使用本文提出的加权多源的TrAdaBoost回归算法;

(3) Muti-TrAdaBoost without γ表示不使用本文提出的误差容忍系数,在实验中误差容忍系数γ取值为0.05(通过交叉验证方式得到的最优参数)。

本文考虑使用拟合度R^2和均方误差MSE两个指标,综合衡量回归模型的效果,表3展示的结果为每组实验的平均值。实验验证了加权多源的TrAdaBoost回归算法的有效性,误差容忍系数可以提高大约0.01的R^2分数。

3.2 异常站点检测与真值预测实验结果

本文使用了10个省份的数据,其中A,B,C三省的站点数量最少,分别为267,340,471,将作为目标领域,其他7个省份作为源领域。首先,在不使用迁移学习的情况下,单独训练每个省份的数据,取R^2平均值以选取最合适的基回归器,实验结果(见图4)发现SVR最适合本任务。

本文考虑以下几种对比实验:Target表示只用目标省份的数据进行训练;Target+Source表示将目标数据和辅助数据放在一起进行训练;Muti-TrAdaBoost表示使用本文提出的加权多源的TrAdaBoost回归算法进行训练;Muti-TrAdaBoost without γ表示在算法执行中不使用本文提出的误差容忍系数γ。实验结果见表4。通过实验结果可以发现:由于数据量过少,只用目标省份的数据很难训练出满意的模型;迁移学习考虑了目标领域与源领域数据分布的差别,在模型的分数上有一定的提升;容忍系数γ的引入,一定程度上解决了在回归问题中,源领域样本权重缩减太快的问题,从而提高了算法的效果。

召回率可以很好地衡量模型对异常站点的检测能力。因为人工排查所有站点的业务数量耗时耗力,所以本文采用随机抽样的方式,每次抽样100个站点,删减其业务数量,然后观测模型能否检测到被删减的站点。在A,B,C三省的站点中,随机进行了3次无放回抽样删减,平均有93个站点被标记为异常,模型召回率达93%。截至目前,根据模型推荐的异常站点,电网运营人员对约10个异常站点的实际业务数量进行了线下排查,根据排查结果与真值预测结果,得到的R^2分数为0.807 2。

表3 实验结果对比

Tab. 3 Comparison of experimental results

实验编号	AdaBoost-Regressor		Muti-TrAda-Boost		Muti-TrAda-Boost without γ	
	R^2	MSE	R^2	MSE	R^2	MSE
1	0.759	0.248	**0.886**	**0.117**	0.871	0.123
2	0.743	0.240	**0.845**	**0.144**	0.832	0.165
3	0.749	0.255	**0.858**	**0.129**	0.843	0.147

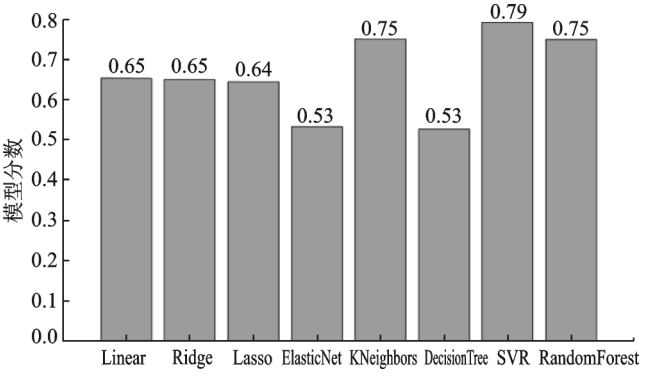

图4 各个回归模型的模型分数对比图

Fig.4 Comparison of model scores for each regression model

表4 算法的结果分析

Tab. 4 Analysis of the results of the algorithm

省份	对比实验	R^2	MSE
A	Target	0.512	396.107
	Target+Source	0.794	141.542
	Muti-TrAdaBoost	**0.825**	**106.329**
	Muti-TrAdaBoost without γ	0.821	110.775
B	Target	0.637	240.273
	Target+Source	0.783	133.995
	Muti-TrAdaBoost	**0.809**	**103.474**
	Muti-TrAdaBoost without γ	0.798	110.248
C	Target	0.657	219.485
	Target+Source	0.803	147.335
	Muti-TrAdaBoost	**0.827**	**132.426**
	Muti-TrAdaBoost without γ	0.819	138.932

4 结束语

本文提出了加权多源TrAdaBoost的回归算法,多源的迁移学习拥有更广泛的使用场景和有效避免

负迁移的优势。同时提出了误差容忍系数,该系数能够一定程度解决源领域样本权重缩减过快的问题,提高了算法的效果。在修改后的Friedman #1回归问题上进行了实验,验证了该算法的有效性。本文将提出的算法应用到电力通信网的行业问题中,提出了异常站点(业务数量缺失严重的站点)检测与真值预测模型,在特征工程中使用了社交网络分析的方法,最终的实验效果进一步验证了算法的有效性。本文将多源迁移学习应用到电力通信网的行业问题中,可以给该行业的问题带来新的思路和方法。

参考文献:

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

[2] Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1): 9.

[3] Davis J, Domingos P. Deep transfer via second-order markov logic[C]//Proceedings of the 26th Annual International Conference on Machine Learning. [S.l.]: ACM, 2009: 217-224.

[4] Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning[C]//Proceedings of the 24th International Conference on Machine Learning. [S.l.]: ACM, 2007: 193-200.

[5] Sugiyama M, Suzuki T, Kanamori T. Density ratio estimation in machine learning[M]. Cambridge: Cambridge University Press, 2012.

[6] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2011, 22(2): 199.

[7] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment[C]//Proceedings of the 19th International Conference on World Wide Web. [S.l.]: ACM, 2010: 751-760.

[8] Zhao Z, Chen Y, Liu J, et al. Cross-people mobile-phone based activity recognition[C]//IJCAI. Barcelona, Catalonia, Spain: [s.n.], 2011: 2545-2550.

[9] Al-Halah Z, Rybok L, Stiefelhagen R. What to transfer? high-level semantics in transfer metric learning for action similarity [C]//Pattern Recognition (ICPR), 2014 22nd International Conference on. [S.l.]: IEEE, 2014: 2775-2780.

[10] Yao Y, Doretto G. Boosting for transfer learning with multiple sources[C]//Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2010: 1855-1862.

[11] Cheng Y, Cao G, Wang X, et al. Weighted multi-source TrAdaBoost[J]. Chinese Journal of Electronics, 2013, 22(3): 505-510.

[12] Tan B, Song Y, Zhong E, et al. Transitive transfer learning[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2015: 1155-1164.

[13] Tan B, Zhang Y, Pan S J, et al. Distant domain transfer learning[C]//AAAI. [S.l.]: AAAI, 2017: 2604-2610.

[14] Ruan Yuncui. Research on the method of improving the basic data quality of transportation management system (TMS)[J]. Digital Communication World, 2016(11): 26, 60.

[15] Liu Fuxin, Li Yisong, Cui Mengyu, et al. Research and application of data quality monitoring system for power big data[J]. Computer Knowledge and Technology, 2016, 12(31): 3-5.

[16] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey[J]. ACM Computing Surveys (CSUR), 2009, 41(3): 15.

[17] Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: A survey[J]. Data Mining and Knowledge Discovery, 2015, 29(3): 626-688.

[18] Bhuyan M H, Bhattacharyya D K, Kalita J K. Network anomaly detection: Methods, systems and tools[J]. IEEE Communications Surveys & Tutorials, 2014, 16(1): 303-336.

作者简介:

杨济海(1983-),男,教授级高级工程师,研究方向:电力信息通信技术,E-mail:jxjihai@139.com。

李号号(1992-),男,硕士生,研究方向:机器学习,E-mail:a981945164@163.com。

彭汐单(1983-),女,高级工程师,研究方向:电力系统大数据及应用,E-mail:39393697@qq.com。

张智成(1979-),男,工程师,研究方向:电力信息化管理咨询、应用系统设计等,E-mail:zhangzhicheng2@sgepri.sgcc.com.cn。

黄倩(1994-),女,硕士研究生,研究方向:大数据,E-mail:qianhwang@qq.com。

李石君(1964-),男,教授,博士生导师,研究方向:大数据,E-mail:shjli@whu.edu.cn。

(编辑:夏道家)