计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):270−278 [doi: 10.15888/j.cnki.csa.008726]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于集成迁移学习的机械钻速预测①

杨顺辉[1], 郭珍珍[2], 张洪宝[1], 高明亮[3]

[1](中国石油化工股份有限公司 石油工程技术研究院, 北京 100101)
[2](西南石油大学 计算机科学学院, 成都 610500)
[3](西北民族大学 电气工程学院, 兰州 730124)
通信作者: 郭珍珍, E-mail: gzz18784188947@163.com

摘　要: 在钻井过程中, 钻速是指机械钻头破岩加深钻口的速度, 是反映钻井效率的一个重要指标. 近年来机器学习方法被应用于机械钻速预测, 然而实践中发现这些方法应用于新油田时, 预测精度显著下降, 主要原因是新油田可供学习训练的数据通常很少甚至完全缺失. 因此提升针对新油田的机械钻速预测性能是一个有待解决的问题. 针对该问题, 本文提出了一种基于迁移学习的跨油田机械钻速预测方法, 构建了一种带物理约束的集成迁移回归模型预测新油田的机械钻速. 在真实钻井数据集上的实验表明, 本文提出的机械钻速预测方法是有效的, 预测精度也显著优于目前主流的同类方法.

关键词: 机械钻速预测; 物理约束; 迁移学习; 集成算法; 跨领域知识

引用格式: 杨顺辉,郭珍珍,张洪宝,高明亮.基于集成迁移学习的机械钻速预测.计算机系统应用,2022,31(10):270–278. http://www.c-s-a.org.cn/1003-3254/8726.html

Rate of Penetration Prediction Using Ensemble Transfer Learning

YANG Shun-Hui[1], GUO Zhen-Zhen[2], ZHANG Hong-Bao[1], GAO Ming-Liang[3]

[1](Research Institute of Petroleum Engineering, China Petroleum and Chemical Co. Ltd., Beijing 100101, China)
[2](School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)
[3](College of Electrical Engineering, Northwest Minzu University, Lanzhou 730124, China)

Abstract: In the process of drilling, the speed at which a drill bit breaks through rock and deepens the drill hole is called the rate of penetration (ROP), which is an important index reflecting drilling efficiency. In recent years, machine learning methods have been applied to the ROP prediction. However, it is found in practice that the prediction accuracy of ROP based on existing machine learning methods is significantly reduced when applied to new oil fields, and the main reason is that the data available for learning and training in these new fields are usually scarce or even completely missing. Therefore, improving the prediction performance of ROP in new oil fields is an important issue to be solved. Considering this, a cross-oilfield ROP prediction method based on transfer learning is proposed, and a boosting transfer regression model with physical constraints is constructed to predict ROP of new oil fields. The experiments on real drilling datasets indicate that the proposed method is effective, and the prediction accuracy is significantly better than that of the current mainstream ROP prediction methods.

Key words: rate of penetration prediction; physical constraint; transfer learning; ensemble algorithm; cross-domain knowledge

近代工业革命后, 能源成为了人类社会生活中赖以生存生活的重要构成部分. 石油资源被称为“工业的血液”[1], 不仅是一种不可再生的资源, 更是国家生存和发展不可或缺的战略资源, 是当今世界各国的经济命

① 基金项目: 国家自然科学基金 (61861038)
收稿时间: 2022-01-07; 修改时间: 2022-01-30; 采用时间: 2022-02-22; csa 在线出版时间: 2022-06-30

脉. 石油的形成过程极其复杂缓慢, 不可再生的石油资源就变得十分关键. 伴随着经济社会的快速发展, 带动着自然资源的消耗也逐年增大, 对石油、天然气等自然资源的使用急剧增长[2]. 全球从陆地到海洋, 从浅层到中层、再到深层的勘探来满足日常的生活需求. 经过长达多年来的石油勘探, 我国在浅层和中层的石油储量已经基本勘探清楚, 剩下不多井正在开采. 然而, 这已远远无法满足社会需要[3]. 同时, 由于实际钻井过程施工情况复杂, 工况变化多样, 获取的录井参数环境呈现出明显的非平稳性, 并且采用人为的方式获取录井参数成本昂贵, 影响因素极多难以考虑完全, 钻井效率受到严重的影响. 因此, 如何提高钻井效率、提升钻井速度是当今国内外研究的热点课题[4].

在钻井工程中, 钻头钻破岩石加深钻孔的速度称为机械钻速. 机械钻速是反映钻井效率的一个关键指标, 受到钻头尺寸、钻井参数、岩石岩性等诸多因素的影响和制约[5], 它与开采成本、开采时间有着直接关联[4]. 钻速预测对于钻井参数的确定和钻井成本的优化是必要的. 钻井机械钻速的准确预测, 能够有效地估算钻井周期, 从而根据预测结果优化配置资源, 可以减少钻井开采成本、增大石油产量, 这对于企业降低钻井施工成本、减少钻井风险, 对于国家能够解决能源紧缺问题等有着重大意义.

随着大数据技术的飞速发展以及数据规模的急速增长, 采用机器学习的方法对数据进行挖掘并应用到钻井过程当中, 与基于物理模型的方法相比, 机械钻速的预测精度有着显著的提高. 传统的机器学习方法通常建立于数据独立同分布这一假设之上[6], 然而在实际钻井过程中, 不同的油田信息具有明显不同的模式, 现有的机器学习方法使用已钻井数据预测新油田时, 预测精度显著下降, 如何高效地进行机械钻速预测并将其应用于后续各种油田处理在石油领域中面临着长期的挑战. 优秀的网络模型皆是基于大量标注数据集 (如 COCO、ImageNet) 训练得到, 然而实际应用中高质量且具有标签的大型井下数据集资源匮乏, 难以支撑优秀网络模型, 可能产生严重的过拟合问题. 迁移学习不受训练数据集与目标数据之间关系的约束, 能够根据不同任务之间的相似性, 实现源域的已有知识迁移, 可有效解决过拟合问题. 目前, 迁移学习方法已经在钻井工程中的岩性识别、钻头选择、异常工况检测等多种场景得到了广泛的应用[7]. 针对钻井过程中机械钻速预测这一回归问题, 本文以真实历史钻井数据钻头尺寸、钻压等字段为特征, 以机械钻速为标签, 采用将迁移学习与物理模型相结合的方法, 提出一种基于集成迁移学习的机械钻速预测模型. 实践中, 采用真实钻井数据, 尝试了包括 linear regression (线性回归)[8]、传统的 AdaBoost 回归、只有目标域数据进行训练和几种先进的基于特征与基于实例的迁移学习方法[9] 等建模方法, 采用多种回归评价指标衡量模型的性能, 证明了本文提出的方法进行跨领域机械钻速预测的有效性, 钻速预测精度也得到显著提高.

1 相关工作

在钻井过程中, 提速提效是永恒不变的追求目标. 机械钻速 (ROP) 的准确预测可显著缩短钻井作业时间, 节约钻井成本. 机械钻速受到多种因素的影响和制约, 有可控因素和不可控因素[10]. 可控因素是指通过一定的设备和技术手段可进行人为调节的因素, 如地面机泵设备、钻头尺寸、钻井液性质、钻压、转速. 不可控因素是指客观存在的因素, 如所钻的地层岩性、储层埋藏深度以及地层压力等. 针对机械钻速预测, 其研究进展大体可以分为 3 个阶段: 用现场数据直接统计出钻速方程, 考虑所钻地层性质和钻头结构的钻速方程, 用计算机仿真方法来预测机械钻速.

1.1 传统方法

国内外学者都提出了各自与地层特性和钻头结构性质相关的钻速方程. 1974 年, Bourgoyne 等[11] 将机械钻速视为钻头压力、转速等 8 个参数的函数, 但该方法存在局限性, 只适用于牙轮钻头情况. 2008 年, Rastegar 等[12] 在前人的基础上提出改进的 ROP 预测模型, 同时考虑了钻头水力参数、钻头的磨损情况和岩石强度等因素的影响. 传统的物理建模方法给机械钻速预测带来了可见的影响, 但方法大多根据专业知识经验, 建模方法高度依赖于岩石岩性, 模型泛化性能不佳. 且由于校准需要进行不断变化, 从而限制了其函数的形式. 随着大数据和机器学习的迅速发展, 很多学者开始将机器学习方法应用到机械钻速预测方面. 2004 年, Rommetveit 等[13] 提出了一种新型的钻井自动化模拟系统, 通过对比实测数据和预测数据得到钻井过程中的实时诊断结果, 但是该系统还处在功能设想阶段, 目前尚未实现全部功能, 且考虑的 ROP 影响因素较少; 2008 年, Bahari 等[14] 基于文献 [11] 提出的模

型并结合遗传算法计算了机械钻速预测模型参数, 但该研究只对 ROP 进行了计算预测, 并没有作进一步的优化分析. 在数据量较充足、数据质量较高的条件下, 采用多元回归[15] 等机器学习方法构建的预测模型的预测准确度较高, 能够在当前设备和资源条件下准确找寻影响机械钻速的若干个核心因素.

传统的机器学习方法大多借助监督学习的推动, 依赖于已有数据, 即需要足够多的标注好的训练样本进行学习, 在数据样本稀少的场景下, 性能会显著下降. 对新领域执行机器学习常遇到标注稀缺问题, 获取大量标注数据成本较高且耗时, 严重制约了经典监督学习方法的效果. 同时, 伴随着多领域、多媒体大数据的不断涌现, 如何研究自动方法对其进行跨领域分类和组织变得愈加重要[16]. 在机器学习的领域中, 已经开发了许多用于迁移学习的方法, 通过将在源数据上的预训练模型迁移到感兴趣的目标数据上, 迁移学习思想被证明是更具有普遍有用的. 迁移学习放宽了经典监督学习中关于训练数据和测试数据服从独立同分布这一基本假设, 将相似但具有不同分布的源域和目标域数据映射到同一个特征空间, 尽可能地保留映射后数据的属性同时缩小数据的维度, 最小化两个领域的概率分布差异. 当源域和目标域数据来自不同的分布时, 通常采用领域分布自适应 (domain adaptive, DA) 算法[17] 来弥补分布差异.

1.2 深度学习

近年来, 深度学习方法在计算机视觉中取得了令人瞩目的成功. 刘胜娃等[18] 结合人工神经网络技术领域知识, 提出一种基于人工神经网络的定向井机械钻速预测模型, 该模型在数据量充足的情况下, 预测准确性较高. 文献 [19] 通过建立渤中区域深层机械钻速预测神经网络模型, 能够在当前特定区域条件下准确找寻影响机械钻速的若干个核心因素. 目前使用的深度网络模型假设训练数据和测试数据为相同的分布, 然而在实际钻井过程中, 训练数据和测试数据的分布往往并不相同, 高质量且具有标签的大型井下数据集资源匮乏, 难以支撑优秀的深度网络模型, 这导致训练得到的模型鲁棒性能较差. 迁移学习不受源域数据与目标数据之间关系的约束[9], 对于缺乏标记数据的目标任务, 有很强的动机来构建有效的学习者, 利用来自相关源域的丰富标记数据, 将已训练好的模型参数迁移到新模型进行训练. 研究表明, 先前对象的认识与新对象的相似性和联系, 有助于新对象的学习. 在特定数据集或任务上训练的 CNN 模型可以针对不同领域的新任务进行微调.

随着深度学习在各个领域的广泛应用, 大量的深度迁移学习[20] 方法被提出. 深度迁移学习 (deep transfer learning, DTL) 通过将深度学习与迁移学习相结合, 将辅助领域训练的深度模型重用于目标领域, 能够有效地降低模型的训练时间, 使现有数据得到更充分的利用, 提高深度网络在实际应用中的泛化能力. 对比传统的非深度迁移学习方法, 深度迁移学习方法在不同的学习任务上得到一定的提升. 神经网络体系结构基于丰富标记的源域数据和标注缺失的目标域数据进行训练, 根据目标任务进行结构调整, 经过目标数据的再次训练, 形成最终的目标网络, 能够有效地促进特征的出现. 若此目标网络优于未经迁移的网络, 则该迁移为正迁移, 反之则为负迁移[7].

2 基于集成迁移学习的机械钻速预测方法

2.1 问题定义

在迁移学习当中, 包含两个基本的概念, 分别是领域 (domain) 和任务 (task). 领域D是进行知识学习的主体, 主要有数据以及生成这些数据的概率分布 P 所组成[21]. 在迁移学习中对应两个基本的领域, 分别是源领域 (source domain, D_s) 和目标领域 (target domain, D_T). 源领域是指有知识、有丰富数据标注的领域, 属于迁移对象. 目标领域就是需要最终赋予知识的对象, 一般来说, 目标领域当中大部分都是未标注数据. 任务 T 指的是学习的目标, 由标签和标签对应的函数组成. 迁移学习旨在从一个或多个源领域中提取知识, 并将知识应用于目标任务当中.

给定一个有标签的源域数据$D_s = \{x_i, y_i\}_{i=1}^n$和一个无标签的目标域$D_T = \{x_j\}_{j=1}^m$. 两个领域的数据概率分布$P(x_s)$和$P(x_t)$不同, 即$P(x_s) \neq P(x_t)$. 迁移学习的目标就是要借助源域$D_S$的先验知识来学习目标领域$D_T$的知识 (标签)[22].

假定源域和目标域的特征空间和样本空间分别相同, 即$X_S = X_T$且$Y_S = Y_T$, 但两个领域的特征分布不同, 即存在条件概率分布不同$Q_S(y_s|x_s) \neq Q_T(y_t|x_t)$或者边缘分布不同$P_S(x_s) \neq P_T(x_t)$. 领域自适应就是源域和目标域不一样, 具体来说, 两个领域的数据概率分布不同, 但是两个领域共享相同的特征和类别, 其维度是一致的[17]. 此刻, 迁移学习的目标就是利用有标记的数据来学习一个分类器f来预测目标领域x_t.

2.2 钻前机械钻速预测模型简介

集成学习是通过将许多弱分类器进行集成提升为强学习器的过程[23]. 一般来说, 用得比较多的是同质学习器, 即同质集成中的个体学习器属于同种类型. 同质学习器根据基学习器之间是否存在依赖关系分为 Boosting 系列算法[24] 和随机森林系列算法. AdaBoost 作为提升算法 (Boosting) 的一种, 根据基学习器的学习误差率来更新训练样本的权重值, 增加学习误差率高的训练样本权重, 再基于调整样本权重后的训练集训练基学习器, 不断调整基学习器的权重, 将这些弱学习器进行线性组合形成一个强学习器, 进而达到提升整体准确率的效果. 算法的性能通过“少数服从多数”这一方法进行投票决出结果. 随着集成中基学习器数目的不断增加, 集成的错误率将指数级下降, 最终将趋于 0.

基于 Boosting 的迁移学习算法, 也称为 TrAdaBoost 算法, 是由 Dai 等[25] 提出的一种典型迁移学习算法. TrAdaBoost 算法假设源领域和目标域数据具有完全相同的特征与标签空间, 但两者的数据分布不同. 将源域数据与部分目标域数据整合得到训练实例, 由于源域与目标域之间的分布差异, 源域数据样本可能会对目标任务的学习有利, 有可能没有, 甚至有可能有害. TrAdaBoost 算法通过对训练实例赋予权重, 增加被错误分类的目标实例的相对权重[26]. 当源实例被错误分类时, 降低其权重值, 具体来说, 就是给数据乘上一个 0 到 1 的值. 在下一次分类, 被错误分类的样本对分类模型的影响就会比上一次迭代小一些. 通过这种方式, TrAdaBoost 旨在识别和利用与目标数据最相似的源实例, 而忽略那些不相似的实例. TrAdaBoost 算法模型具体描述如图 1.

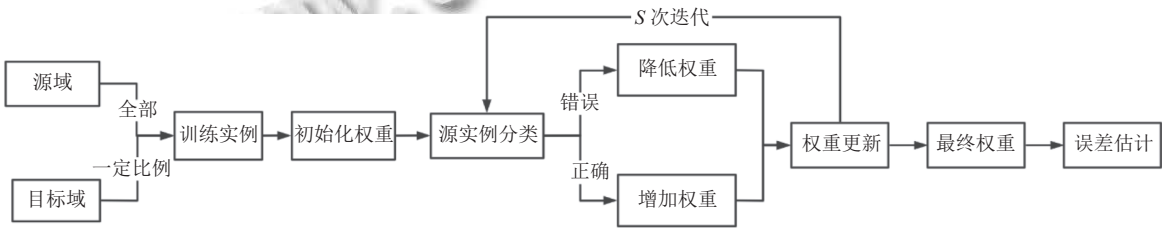

图 1 TrAdaBoost 模型架构图

2.3 带物理模型约束的集成机械钻速回归模型

将钻井数据看作连续的, 在统计学上称为回归问题. 结合 TrAdaBoost 的原理与传统的回归算法产生了新的回归算法 TrAdaBoostR2[27]. 作为一种基于样本实例的迁移学习方法, TrAdaBoostR2 对每个训练实例进行加权, 确保迁移的源域知识与目标任务相关. 然而, 当源域样本数远大于目标域时, 目标实例的总权重可能需要多次迭代才能接近源实例的总权重, 此时目标数据的权重可能会严重偏斜, 那些异常值或与源数据最不相似的目标实例可能会代表大部分权重[28]. 其次, 即使是那些代表目标概念的源实例, 它们的权重最终也趋于零.

基于 Bingham (1965) 提出的基本 ROP 模型[29], 已知转速 (ROP)、钻头压力 (RPM) 和钻头直径 (D_b) 等参数, 可以通过式 (1) 计算得到机械钻速的预测值. 其中, α和γ为岩性模型的经验参数:

$$ROP = \alpha RPM\left(\frac{WOB}{D_b}\right)^{\gamma} \tag{1}$$

物理模型虽然源自钻井物理原理, 但是涉及经验参数和拟合函数的约束, 这常常会导致较差的结果. 为了克服上述缺点, 选择基于样本实例迁移方法 TrAdaBoost.R2 作为基础, 提出一种带物理模型约束的集成迁移学习方法, 分两个阶段对样本实例进行调整. 算法在第一阶段, 源实例的权重逐渐向下调整, 直到达到某一个值 (该值采用交叉验证确定得到); 在第二阶段, 首先对所有源实例的权重冻结, 而目标实例的权重在 AdaBoost.R2 中被正常更新, 只有在第 2 阶段生成的假设被存储并用于确定结果模型的输出.

假定存在n个源域训练数据$D_{S_1},\cdots,D_{S_n}$, m个用于训练的目标域数据$D_{T_1},\cdots,D_{T_m}$, 迁移学习的目的就是充分利用有标记的源域数据来提高目标分类器f_T的学习效率. 首先, 定义第 h 次迭代训练实例的权重向量 $w^h=(w^h_{S_1},\cdots,w^h_{S_n},w^h_T)$, 其中, w_S表示源域数据样本实例的权重, $w_T=(w^1_T,\cdots,w^m_T)$表示目标域数据用来训练的m个样本权重向量. 初始化权重为:

$$w^1_i=\frac{1}{n+m},1\leqslant i\leqslant n+m \tag{2}$$

清空候选基学习器并对现有的权重进行规范化, 选择基学习器f_t对训练集$D_{\text{train}}=D_S\cup D_{T-\text{train}}$进行训练. 为了保证模型不会因为目标实例划分成训练集和

测试集而造成误差, 采用十折交叉验证. 将目标领域数据集随机划分 10 份, 随机选择其中一份作为测试集, 剩下的 9 份与源实例进行整合作为训练集进行实验, 依次进行 10 组实验. 同时, 采用 Bingham 提出的基本 ROP 模型对算法进行物理约束, 采用式 (3) 计算基学习器 f_t 在 $D_{T-\text{test}}$ 上的误差值, 选取平均绝对误差最小的用于后续模型.

$$(e_t^i)^k=\sum_{j=1}^{m}\frac{w_t^j[|(f_t^i)^k-y_t^j|+|(y_t^p)^j-y_t^j|]}{\sum_{i=1}^{m}w_t^i} \tag{3}$$

其中, $(y_t^p)^j$ 表示采用物理模型 (即式 (1)) 计算得到的 *ROP* 值, y_t^j 表示目标域数据的真实标签值, $(f_t^i)^k$ 表示第 i 个基分类器进行 k 折交叉验证预测得到的 *ROP* 值. 根据误差估计来更新训练样本实例的权重值, 误差越大, 其权重设置越小. 对其进行 S 次迭代, 并对权重进行更新, Z_t 是一个归一化的常量, 使得最终目标实例的权重为 $\frac{m}{n+m}+\frac{t}{(S-1)}(1-\frac{m}{n+m})$.

$$w_i^{t+1}=\begin{cases}w_i^t\beta_t^{e_i^t}/Z_t, 1\leqslant i\leqslant n\\ w_i^t/Z_t, n+1\leqslant i\leqslant n+m\end{cases} \tag{4}$$

对源实例的权重更新采用加权多数算法 (即 WMA[27]) 机制, 第 2 阶段首先对所有源实例的权重冻结, 采用 Bootstrap 对观测信息进行多次重复抽样, 建立起充足的样本, 采用基学习器对取样的样本进行预测, 并计算损失函数, 采用 TrAdaBoostR2 来更新目标实例的权重向量, 最后对权重进行规范化处理, 生成的模型被存储并用于确定结果模型的输出.

本文选用决策树回归 (decision tree regressor) 算法[30] 作为基学习器进行集成迁移回归, 对模型参数进行调整, 不断更新模型权重. 在模型优化问题中, 通过计算真实值与预测值的平均绝对误差 (mean absolute error, *MAE*) 作为模型性能的一个衡量指标. 平均绝对误差作为回归损失函数中常用的误差计算, 通过计算预测值与真实值之间差值绝对值和的均值, 可以有效地避免误差相互抵消, 因而可以较准确地反应实际预测误差的大小. 其中, y_{pred} 表示模型最终的预测值, y_i 表示相应的实际值.

$$MAE=\frac{\sum_{i=1}^{n}|y_i-y_{\text{pred}}|}{n} \tag{5}$$

本文提出了一种带物理模型约束的集成迁移回归模型来对钻前机械钻速进行预测, 算法具体描述如算法 1.

算法 1. 基于集成迁移回归的机械钻速预测算法

输入: D_S: 源域数据集; D_T: 目标域数据集
输出: *err*: 平均绝对误差 *MAE* 值

1. 初始化源域数据集 $D_s=[n_s\times m_s]$, 目标域数据集 $D_T=[n_T\times m_T]$;
2. 确定基学习器 f_T 为决策树回归算法;
3. $feature=D[:,1:m-1]$; $label=D[:,-1]$
4. 采用 One-Hot 和 Z-score 标准化对数据集进行预处理;
5. 将目标域划分为训练集和测试集: D_{train} 和 D_{test};
6. 确定最大估计次数 $n_{\text{estimator}}$, 步骤数 S, K=10;
7. 初始化权重为 $W_i^1=1/(n+m)$;
8. **for** $i=1\rightarrow S$ **do**
9. $D\leftarrow(D_{\text{train}}+D_s)$
10. 清空候选基学习器, 对现有权重进行规范化;
11. 采用 TrAdaBoostR2 进行训练得到模型model;
12. **for** $j=1\rightarrow K$ **do**
13. $D_{\text{train}}^j=(D_S+D_{T-\text{train}})$
14. //对用于训练的目标实例进行权重更新
15. 采用 TrAdaBoostR2 对 D_{train}^j 进行训练并计算 $D_{T-\text{test}}$ 的预测值;
16. 采用式 (3) 计算误差估计;
17. 采用式 (4) 更新权重 $W_i^{(j+1)}$;
18. 确保目标实例总权重不随交叉分割而改变;
19. **end for**
20. 采用 Bootstrap 对 X 进行多次重复采样;
21. 使用基学习器更新目标实例的权重向量;
22. 返回样本权重 W_i;
23. **end for**
24. 得到最优的样本权重 W^*;
25. 计算目标域数据的预测值 $y_{\text{pred}}=\text{model}(f_T,W^*)$;
26. 采用式 (5) 计算预测值与真实值的误差值 *err*;
27. 返回 *err*.

3 实验分析

3.1 实验设置

本论文采用的数据集共包括 156 次测量, 这些测量是从特定区块的 26 口 S 井和 3 口 WD 井收集得到的历史钻井数据. 实验数据具体描述如表 1, 每个样本数据包含斜深 (depth)、钻压 (wob)、大钩载荷 (hook_load)、泵压 (spp)、转盘转速 (bit_rpm)、泵排量 (flow_rate)、扭矩 (torque)、地层类型 (formation)、钻头类型 (bit_type)、钻头尺寸 (bit_size)、岩性类型 (lith) 等 51 个特征参数和 1 个机械钻速 (ROP) 样本标签. 通过对数据进行预处理操作, 有效保留数据样本在各个维度上的信息分布. 同一口井有多次测量, 其测量结果是连续的, 为了保证井口数据的完整性和独立性, 将同一钻井数据作为一个整体, 目标域共有 3 口井数据, 分别是 WD1、WD2 和 WD3, 采用交叉验证进行模型预测, 通过随机选择目标域数据将其与源域数据整

合作为训练集, 剩下的钻井数据作为测试集样本.

本文使用 51 个特征参数作为机械钻速预测模型的输入, 由于数据集中的字符类型特征无法被机器模型学习, 因此在建模时需要将其转化成易于机器利用的数值型特征. 独热编码 (one-hot) 用 N 位状态寄存器来对 N 个状态进行编码, 从而将类别变量转换为数值变量, 由于 one-hot 编码后的特征值只有 0 或 1, 因此采用该方法不会影响原类别特征在模型中的权重比例. 采用独热编码进行数据类型转换会将数据维度扩大, 为了进一步排除数据集维度扩大对实验的影响, 再对其采用主成分分析[31] 对数据维度进行降维. 同时, 利用原始数据的均值和标准差对其进行标准化, 使处理后的数据符合标准正态分布. 数据的标准化处理能够有效地提升模型精度, 加快训练网络的收敛性. Z-Score 标准化对样本数据在不同特征维度进行伸缩变换, 使得不同度量之间的特征具有可比性, 并且不会改变原始数据的分布, 通过将不同量级的数据转化为统一量级的 Z-Score 分值进行比较, 能够在特征提取时有效保留样本各维度上的信息分布. Z-Score 标准化的数学表达如下, 其中μ, σ表示原始数据的均值和标准差.

$$x = \frac{x-\mu}{\sigma} \tag{6}$$

采用决策树回归模型作为该模型的弱学习器, 同时对模型参数进行调整, 得到具体参数设置如表 2, 其中, n_estimators 表示的是迭代次数, 也就是本次实验中采用弱学习器的个数; learning_rate 表示学习率; steps 表示的是步骤数; folds 表示的是交叉验证的折叠次数; max_depth 表示的是每一个学习器的最大深度, 用于限制回归树的节点数目.

表 1 实验数据介绍

数据集	类型	测量次数	样本规模
S01-S26	Source	148	(23 386, 51)
WD	Target	8	(1877, 51)
WD1	Target	4	(680, 51)
WD2	Target	2	(511, 51)
WD3	Target	2	(686, 51)

表 2 模型参数的选择

参数	设置
n_estimators	100
learning_rate	0.5
steps	30
folds	5
max_depth	7

3.2 结果分析

对预处理后的数据进行模型训练与验证. 根据领域特点, 采用交叉验证对模型进行训练与预测, 通过对目标域 WD 数据中随机选择一份作为测试集, 剩下两口井与源域数据整合用于模型训练, 使用线性回归、物理模型作为基线方法, 同时, 为了验证该模型的有效性并保证实验的严谨性, 采用传统的 AdaBoostR2 (即没有采用迁移学习) 模型与基于集成迁移学习的机械钻速预测方法进行实验对比. 通过对模型参数的不断调节, 得到 WD1、WD2 和 WD3 作为测试集下基于集成迁移学习的机械钻速方法的 MAE 值分别为 1.476 04、0.826 26 和 0.857 51. 本文模型算法的性能在表 3 以数字方式进行描述, 在图 2 中以图形方式展示.

表 3 该模型算法的性能对比

源域数据	目标域 (训练集)	目标域 (测试集)	线性回归	物理模型	AdaBoostR2	本文方法
S01-S26	WD2+WD3	WD1	10.159 66	9.122 83	2.353 40	**1.476 04**
S01-S26	WD1+WD3	WD2	2.166 20	3.278 47	**0.749 15**	0.826 26
S01-S26	WD1+WD2	WD3	5.069 46	7.625 03	0.897 71	**0.857 51**

以 WD3 作为测试集为例, 得到该实验设置下线性回归、物理模型、AdaBoostR2 和本文提出算法预测值与真实 ROP 值的对比图. 很明显, 从图 2 中可以看到, 基于集成迁移学习的机械钻速模型大大降低了模型的误差值. 与传统的 AdaBoostR2 方法相比, 基于集成迁移学习的机械钻速预测方法在对峰值进行预测时更接近真实值. 具体来说, 本文方法在 3 种实验设置下的性能分别提升 0.877 36、−0.077 11 和 0.040 2. 同时, 该模型下的 MSE 值也得到了提升, 较传统的 AdaBoostR2 模型 MSE 值降低了 3.003 6.

3.3 对比实验

对实验数据进行同类型操作处理, 设计并进行对比实验. 本文选择了 6 种先进的机器学习方法用于验证本模型方法的有效性. 领域自适应 (DA) 方法通过在一个领域上学习的知识迁移到另一个领域上, 自适应方法分为基于特征的自适应、基于实例的自适应和基于参数的自适应方法. 本文分别选用 3 种基于实例的方法 (KMM[32]、KLIEP[33]、TrAdaBoostR2[27]) 和 3 种

基于特征迁移的方法 (DANN[34]、DeepCORAL[35]、MDD[36]) 用于进行实验对比. 同时, 添加直接采用目标域数据进行训练 (即 TgtOnly) 作为基线方法.

在实验设置上, 为了保证井口数据的完整性和独立性, 以井为单位选取部分目标域数据与源域数据合并一起进行训练, 剩下的目标域作为测试集, 这样达到交叉验证的效果. 将本文提出的基于集成迁移学习的机械钻速预测方法与其他方法进行性能对比, 得到实验结果如表 4. 采用 TgtOnly 分别对 WD1、WD2、WD3 进行机械钻速预测, 计算预测值与真实值的最大均值误差 MAE 值分别为 9.341 4、3.739 9 和 6.632 3. 领域自适应方法通过将一个领域上学习的知识迁移到另一个领域上, 其性能远远好于 TgtOnly. 同时, 采用传统的 TrAdaBoostR2 进行对比验证, 得到本文提出的模型效果明显改善. 实验表明, 本文提出的基于集成迁移学习的机械钻速预测模型拟合效果最佳, 算法的性能远远好于其他几种先进的领域自适应方法, 具有较好的鲁棒性能. 在以 WD2 作为测试集中, MAE 值减小到 0.8263, 相较于这里面最优的方法 KMM 误差减小了 1.387 2, 性能提升了 1.68 倍.

在以 WD2 为测试集中, 得到本文方法与其他几种主流机械钻速预测方法在目标域数据预测的机械钻速值与真实 ROP 的对比图 (图 3), 从图中可以清晰地看到, 本文算法预测得到的机械钻速值与真实标签值具有良好的一致性, 误差值远远小于其他几种主流的机械钻速预测方法, 能够为模型提供相对稳定的效果.

为了更直观进行实验对比, 分别计算本文方法与其他几种机械钻速预测方法的决定系数 (R^2)、均方根误差 (RMSE)、均方根相对误差 (RMSRE) 和平均绝对百分比误差 (MAPE), 多种回归评价指标的对比结果如图 4 所示. 图 4 的结果表明, 本文方法的 R^2 值为 0.868 6, RMSE 值为 0.999 7, RMSRE 值为 0.291 1, MAPE 值为 17.67%, 在多种评价指标中性能最优. 同时图 4 也显示结合迁移学习的机械钻速预测方法相较于不使用迁移学习的方法 (TgtOnly) 性能有明显的提升.

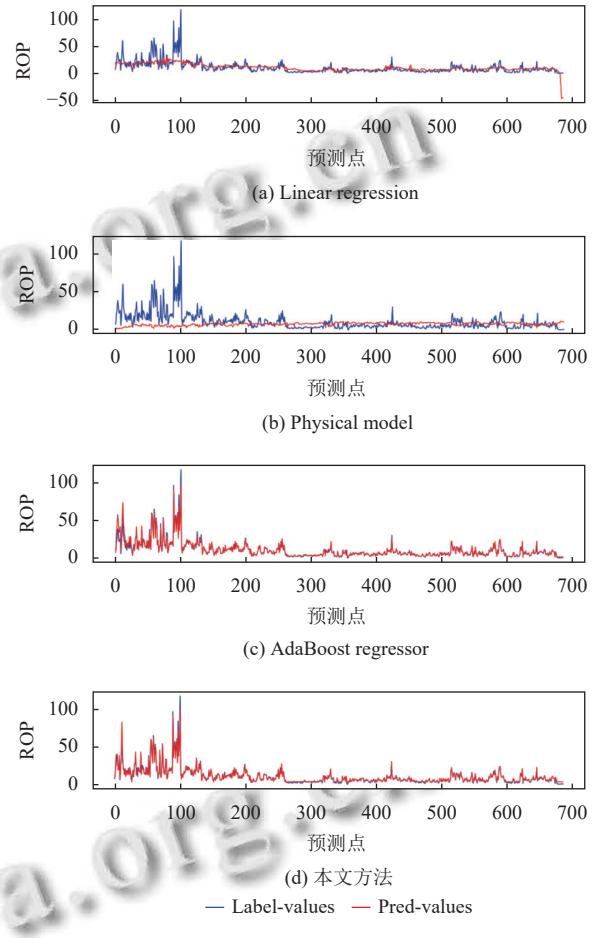

图 2 WD3 做测试集下预测值与真实值对比图

表 4 本文方法与其他方法的性能对比

实验方法	WD1作测试集	WD2作测试集	WD3作测试集
TgtOnly	9.341 399 317 9	3.739 938 767 7	6.632 281 298 1
DANN	8.008 969 745 0	2.472 864 395 3	5.031 593 132 07
DeepCORAL	7.546 214 680 0	2.836 802 762 0	4.736 062 568 62
MDD	7.957 003 423 5	2.492 225 423 3	4.844 078 871 94
KLIEP	7.872 917 797 39	2.293 340 659 3	5.069 057 347 49
KMM	8.385 526 975 86	2.213 431 016 5	6.255 698 774 32
TrAdaBoostR2	6.472 491 732 19	2.818 506 168 4	5.280 384 409 22
本文方法	**1.476 038 799**	**0.826 263 165 9**	**0.857 514 009**

4 结论与展望

本文提出了一种结合物理模型和迁移学习的钻前机械钻速预测方法, 能够在目标油田数据样本缺失或标注的数据样本较少的情形下更准确地预测机械钻速. 本文方法结合机械钻速物理模型, 通过迁移学习识别并利用与目标数据相似的源实例, 确保迁移的知识与目标任务相关. 实验表明, 本文方法机械钻速预测值与实际值之间具有良好的一致性, 与几种主流

机械钻速预测方法相比, 在多种回归评价指标中性能最优. 钻前机械钻速的准确预测能给施工现场提供高效有力指导依据, 为进一步有效地提高钻井效率提供可靠的保障.

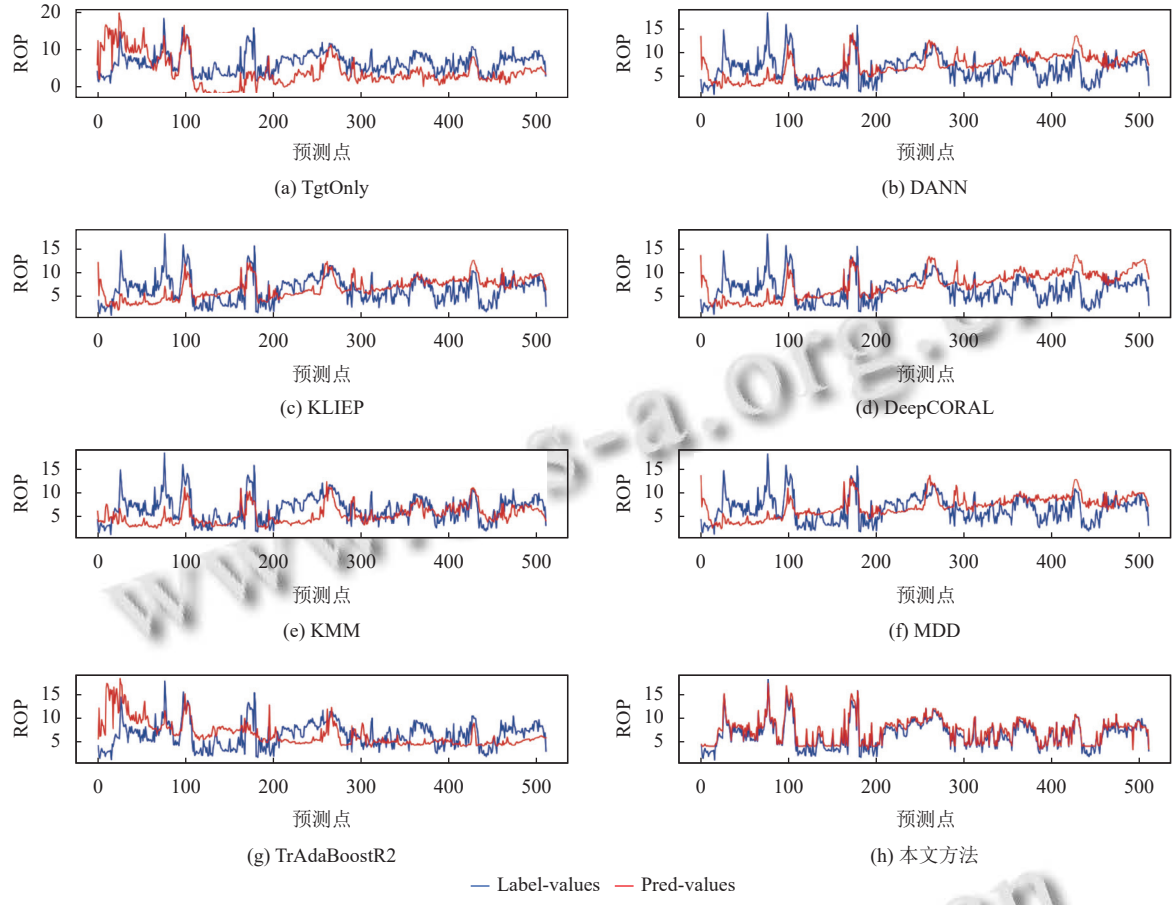

图 3 本文方法与其他机械钻速预测方法在 WD2 上预测值与真实值对比图

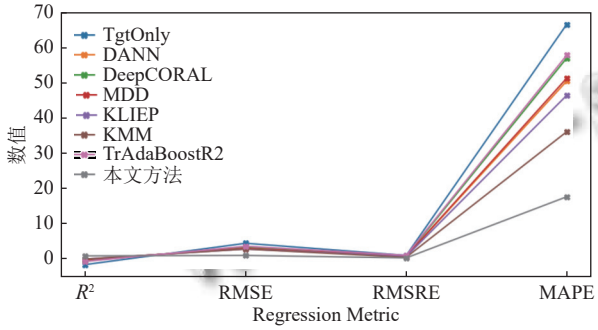

图 4 WD2 数据集上各种方法多种回归评价指标图

参考文献

1 侯明扬. 2020 年全球油气资源并购市场特点及前景展望. 国际石油经济, 2021, 29(3): 45–52. [doi: 10.3969/j.issn.1004-7298.2021.03.008]

2 李振宇, 卢红, 任文坡, 等. 我国未来石油消费发展趋势分析. 化工进展, 2016, 35(6): 1739–1747.

3 顾乐民. 中国石油产量历史回顾与未来趋势. 石油学报, 2016, 37(2): 280–288. [doi: 10.7623/syxb201602016]

4 Shad HIA, Sereshki F, Ataei M, *et al*. Prediction of rotary drilling penetration rate in iron ore oxides using rock engineering system. International Journal of Mining Science and Technology, 2018, 28(3): 407–413. [doi: 10.1016/j.ijmst.2018.04.004]

5 郑红帝. 钻井工程中影响机械钻速的因素分析及应对措施. 石化技术, 2019, 26(1): 248–249. [doi: 10.3969/j.issn.1006-0235.2019.01.170]

6 宋尧. 基于机器学习回归模型的房价预测研究. 电子制作, 2021, (2): 41–43. [doi: 10.3969/j.issn.1006-5059.2021.02.017]

7 吴彦文, 李斌, 孙晨辉, 等. 基于迁移学习的领域自适应推荐方法研究. 计算机工程与应用, 2019, 55(13): 59–65. [doi: 10.3778/j.issn.1002-8331.1810-0199]

8 Amato U, Antoniadis A, de Feis I, *et al*. Penalised robust estimators for sparse and high-dimensional linear models. Statistical Methods & Applications, 2021, 30(1): 1–48.

9 Weiss KR, Khoshgoftaar TM. Comparing transfer learning and traditional learning under domain class imbalance. 2017 16th IEEE International Conference on Machine Learning

and Applications (ICMLA). Cancun: IEEE, 2017. 337–343.
10 Sui D, Nybø R, Azizi V. Real-time optimization of rate of penetration during drilling operation. Proceedings of the 10th IEEE International Conference on Control and Automation. Hangzhou: IEEE, 2013. 357–362.
11 Bourgoyne AT, Young FS. A multiple regression approach to optimal drilling and abnormal pressure detection. Society of Petroleum Engineers Journal, 1974, 14(4): 371–384. [doi: 10.2118/4238-PA]
12 Rastegar M, Hareland G, Nygaard R, *et al*. Optimization of multiple bit runs based on ROP models and cost equation: A new methodology applied for one of the Persian Gulf carbonate fields. IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition. Jakarta: SPE, 2008. SPE-114665-MS.
13 Rommetveit R, Bjørkevoll KS, Halsey GW, *et al*. Drilltronics: An integrated system for real-time optimization of the drilling process. IADC/SPE Drilling Conference. Dallas: SPE, 2004. SPE-87124-MS.
14 Bahari A, Seyed AB. Trust-region approach to find constants of Bourgoyne and young penetration rate model in Khangiran Iranian gas field. Latin American & Caribbean Petroleum Engineering Conference. Buenos Aires: Society of Petroleum Engineers, 2007. SPE-107520-MS.
15 李昌盛. 基于多元回归分析的钻速预测方法研究. 科学技术与工程, 2013, 13(7): 1740–1744. [doi: 10.3969/j.issn.1671-1815.2013.07.006]
16 李晶晶, 孟利超, 张可, 等. 领域自适应研究综述. 计算机工程, 2021, 47(6): 1–13.
17 Lu NN, Chu F, Qi HR, *et al*. A new domain adaption algorithm based on weights adaption from the source domain. IEEJ Transactions on Electrical and Electronic Engineering, 2018, 13(12): 1769–1776. [doi: 10.1002/tee.22739]
18 刘胜娃, 孙俊明, 高翔, 等. 基于人工神经网络的钻井机械钻速预测模型的分析与建立. 计算机科学, 2019, 46(S1): 605–608.
19 王文, 刘小刚, 窦蓬, 等. 基于神经网络的深层机械钻速预测方法. 石油钻采工艺, 2018, 40(S1): 121–124.
20 Tan CQ, Sun FC, Kong T, *et al*. A survey on deep transfer learning. 27th International Conference on Artificial Neural Networks and Machine Learning-ICANN 2018. Rhodes: Springer, 2018. 270–279.
21 刘鑫鹏, 栾悉道, 谢毓湘, 等. 迁移学习研究和算法综述. 长沙大学学报, 2018, 32(5): 28–31, 36. [doi: 10.3969/j.issn.1008-4681.2018.05.008]
22 Pan SJ, Tsang IW, Kwok JT, *et al*. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 2011, 22(2): 199–210. [doi: 10.1109/TNN.2010.2091281]
23 Habrard A, Peyrache J P, Sebban M. A new boosting algorithm for provably accurate unsupervised domain adaptation. Knowledge and Information Systems, 2016, 47(1): 45–73. [doi: 10.1007/s10115-015-0839-2]
24 Yao Y, Doretto G. Boosting for transfer learning with multiple sources. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 1855–1862.
25 Dai W, Yang Q, Xue G, *et al*. Boosting for transfer learning. ACM International Conference Proceeding Series, 2007, 227: 193–200.
26 He HX, Khoshelham K, Fraser C. A multiclass TrAdaBoost transfer learning algorithm for the classification of mobile Lidar data. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 118–127. [doi: 10.1016/j.isprsjprs.2020.05.010]
27 Zhang Q, Li HG, Zhang Y, *et al*. Instance transfer learning with multisource dynamic TrAdaBoost. The Scientific World Journal, 2014, 2014: 282747.
28 Li LC, Ma C. Transfer regression with data-augmented ensemble learning framework. IOP Conference Series: Earth and Environmental Science, 2019, 252(2): 022095.
29 Bingham G. A new approach to interpreting rock drillability. Technical Manual Reprint, Oil and Gas Journal, 1965, 93.
30 Aguilar-Chinea RM, Rodriguez IC, Expósito C, *et al*. Using a decision tree algorithm to predict the robustness of a transshipment schedule. Procedia Computer Science, 2019, 149: 529–536. [doi: 10.1016/j.procs.2019.01.172]
31 Oh SK, Yoo SH, Pedrycz W. Design of face recognition algorithm using PCA-LDA combined for hybrid data pre-processing and polynomial-based RBF neural networks: Design and its application. Expert Systems with Applications, 2013, 40(5): 1451–1466. [doi: 10.1016/j.eswa.2012.08.046]
32 Huang JY, Smola AJ, Gretton A, *et al*. Correcting sample selection bias by unlabeled data. Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge: ACM, 2006. 601–608.
33 Sugiyama M, Nakajima S, Kashima H, *et al*. Direct importance estimation with model selection and its application to covariate shift adaptation. Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver: ACM, 2007. 1433–1440.
34 Ganin Y, Ustinova E, Ajakan H, *et al*. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 2016, 17(1): 2096–2030.
35 Sun BC, Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation. European Conference on Computer Vision. Amsterdam: Springer, 2016. 443–450.
36 Zhang YC, Liu TL, Long MS, *et al*. Bridging theory and algorithm for domain adaptation. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 7404–7413.

(校对责编: 孙君艳)