Mine Engineering 矿山工程, 2024, 12(3), 395-401
Published Online July 2024 in Hans. https://www.hanspub.org/journal/me
https://doi.org/10.12677/me.2024.123048

Hans汉斯

基于石油钻井机械钻速预测的迁移学习理论研究

郑凯文，王国庆，潘陆祥，郭晓乐

重庆科技大学石油与天然气工程学院，重庆

收稿日期：2024年5月14日；录用日期：2024年6月17日；发布日期：2024年7月15日

摘　要

提高钻速预测模型的迁移性有助于实现对钻井机械钻速高效、精准的预测。经过深度调研，总结了迁移学习与传统机器学习的关系，迁移学习是机器学习范畴内一个重要的研究领域；详细对比了迁移学习模式与传统机器学习模式的差异性，并介绍了各类迁移学习方法的特点。深入研究了基于实例的迁移学习方法，对基于迁移学习理论的机械钻速预测模型进行了可行性分析并完成模型设计。

关键词

迁移学习，机械钻速预，机器学习

Research on Transfer Learning Theory Based on Prediction of Penetration Rate of Oil Drilling Machinery

Kaiwen Zheng, Guoqing Wang, Luxiang Pan, Xiaole Guo

School of Petroleum Engineering, Chongqing University of Science and Technology, Chongqing

Received: May 14th, 2024; accepted: Jun. 17th, 2024; published: Jul. 15th, 2024

Abstract

Improving the mobility of drilling rate prediction models can help achieve efficient and accurate prediction of drilling machinery drilling rate. After in-depth investigation, the relationship between transfer learning and traditional machine learning was summarized. Transfer learning is an important research field in the field of machine learning. The differences between transfer learning models and traditional machine learning models were compared in detail, and various

文章引用: 郑凯文, 王国庆, 潘陆祥, 郭晓乐. 基于石油钻井机械钻速预测的迁移学习理论研究[J]. 矿山工程, 2024, 12(3): 395-401. DOI: 10.12677/me.2024.123048

types of transfer learning characteristics of the method were introduced. The case-based transfer learning method was studied in depth, the feasibility analysis of the mechanical penetration rate prediction model based on the transfer learning theory was conducted, and the model design was completed.

Keywords

Transfer Learning, Mechanical Drilling Speed Prediction, Machine Learning

1. 引言

要研究基于机器学习的钻井机械钻速预测模型的迁移性，首先需要深刻认识迁移学习的定义与方法，将传统的机器学习模式与迁移学习对比理解，有助于更好地走进迁移学习的世界。通过对迁移学习的深度调研，根据钻井机械钻速预测模型的实际应用场景，选择出适合的迁移学习方式，对建立具有迁移性的机械钻速预测模型设计出合理方案。

2. 迁移学习的定义

2.1. 机器学习的定义

迁移学习是机器学习范畴内一个重要的研究领域，机器学习与迁移学习的关系示意图见图 1，所以我们从机器学习入手逐步递进能更好的理解迁移学习。1997 年，Tom Mitchell 教授对机器学习给出了一个通用定义。首先引入了三个概念：经验(E, Experience)、任务(T, Task)和表现评价指标(P, Performance measure)；假设一个计算机程序在任务 T 上的性能表现评价为 P，在这个计算机程序积累了经验 E 时，对任务 T 的表现评价 P 有所提升，这个经验积累的过程就称为机器学习。Goodfellow、Bengio 和 Courville 三位教授在合著的《深度学习》中对机器学习概念有一个更为规范的定义，书中是这样阐述的：“机器学习本质上属于应用统计学，更多地关注如何用计算机统计地估计复杂函数，不太关注这些函数提供置信区间[1]。”

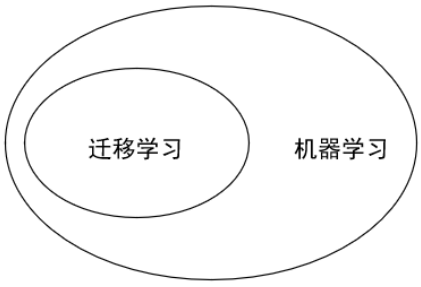

Figure 1. The relationship between machine learning and transfer learning
图 1. 机器学习与迁移学习的关系

首先作如下假设：有一标签空间 X 对应着一个样本空间 Y，则训练数据集可表示为

$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}, x_i\in X, y_i\in Y$；另机器学习的目标函数为 $f\in H$，H 为满足条件的假设空间；则学习目标(目标函数)可表示为：

$$f^*=\arg\min_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}l\left(f\left(x_i\right),y_i\right) \tag{1}$$

其中 $l(\bullet,\bullet)$ 为损失函数。分类问题中常选交叉熵损失(Cross-enteropy loss)为损失函数，回归问题中损失函数通常选最小均方误差(Mean squared error)。

机器学习的目标就是要找到一个最优 f，使得其在训练集中达到最小损失，此时这个目标依据经验风险最小化(Empirical Risk Minimization, ERM)准则，而相应的损失函数就是经验风险[2]。传统机器学习的模式见图 2。

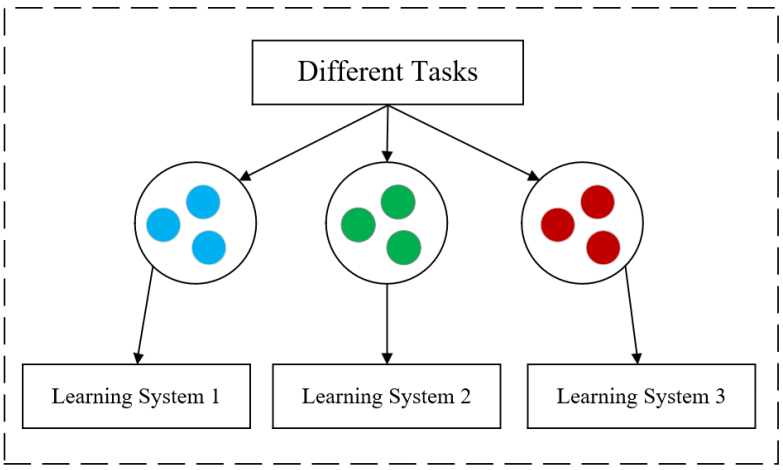

Figure 2. Traditional machine learning model
图 2. 传统机器学习模式

2.2. 迁移学习的定义

迁移学习至少包含两个领域(Domain)：被迁移的领域，称为源领域(Source domain)；待学习的领域，称为目标领域(Target domain)。源域拥有知识也就是大量标注数据，是我们的迁移对象；目标域则是我们要赋予其知识的对象[3]。迁移学习的模式见图 3。

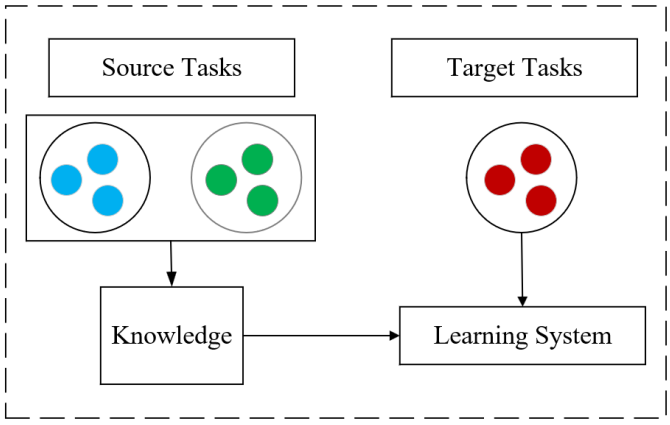

Figure 3. Transfer learning model
图 3. 迁移学习模式

假设特征空间为 X，标签空间为 Y，给定源域 $\mathcal{D}_s=\{x_i,y_i\}_{i=1}^{N_s}$ 和目标域 $\mathcal{D}_t=\{x_j,y_j\}_{j=1}^{N_t}$，其中 $x\in X$，$y\in Y$ 当以下三种情况至少有一种成立时：

① 特征空间不同，$X_s\neq X_t$；

② 标签空间不同，$Y_s\neq Y_t$；

③ 特征空间和标签空间相同，但概率分布不同，$P_s(x,y)\neq P_t(x,y)$。

迁移学习的目标是利用源域数据学习一个预测函数 f，使得模型 f 在目标域上预测误差最小，误差用损失函数衡量，则迁移学习可形式化定义如下：

$$f^*=\arg\min_f E_{(x,y)\in D_t}l\left(f(x),y\right) \tag{2}$$

其中，$E_{(x,y)\in D_t}l\left(f(x),y\right)$表示损失函数在目标域 D_t 上的期望。

3. 迁移学习的分类

将迁移学习根据不同层面分类，其分类结果见图 4，下文主要详细介绍根据学习方式分类的四种具体情况。

3.1. 基于实例的迁移学习方法

基于实例的迁移学习旨在通过从一个或多个相关源任务中学习到的知识，改善在一个目标任务上的性能。这种方法通过将来自源领域的实例(即数据样本点)和其对应的标签用于目标领域的学习过程中，来实现知识的迁移。Dai 等人[4]就提出了基于实例迁移的一个示例算法——TrAdaBoost，它是 AdaBoost 的扩展算法。TrAdaBoost 假设源域与目标域数据具有相同的特征与标签空间，仅是在数据分布上有所不同，并由此认为部分源域数据是有助于目标任务的训练，同时也有另一部分源域数据会对迁移学习的有效性产生负面影响；所以 TrAdaBoost 算法会不断更新源域数据集的权重，来减小“坏”数据的影响，并且提升“好”数据的对目标任务学习的贡献。最后对于迁移性能的评价，仅在目标领域上计算模型预测误差，不再评估模型对于源域数据集的预测效果，这样有利于准确衡量模型在目标任务的表现。

基于实例的迁移学习关键要解决两个问题。一是如何在源域中筛选出与目标域数据有相似分布且具有标签的样本数据；二是如何充分利用这些数据训练模型，使其在目标域上有更高的准确性[5]。

3.2. 基于特征的迁移学习方法

在许多的真实场景中，目标域数据特征往往与源域数据特征不是完全重叠的，相同的特征在源域与目标域中含义却可能不尽相同。甚至会存在一些极端情况，目标域与源域的数据特征根本没有一丝重叠，此时则需要通过特殊方法来转换特征，实现目标域与源域的特征联系[6]。基于特征的迁移学习其核心思想就是我们要把原本不相似的目标域特征空间与源域特征特征空间，通过特征变换学习的方式，抽象到一个共同的特征空间，在此空间中特征服从相同的概率分布。

通常有以下三种方法来实现基于特征的迁移学习方法。一是将域间差异最小化，然后学习目标域及源域的可迁移特征[7]；二是学习目标域与源域都通用的具有高质量的特征；三是通过找到目标域与源域数据的额外相关性，以此拓展特征空间，建立起源域与目标域的特征联系，实现迁移学习。

3.3. 基于模型的迁移学习方法

基于模型的迁移学习也被叫做基于参数的迁移学习，顾名思义，此类方法从源域迁移的主要包括模型参数、模型架构、模型先验知识等模型层次的知识，这样可以直接使用从源域中学习到的模型，省去了对数据重复抽取的过程，也避免了模型对数据进行复杂关系推理的重复过程[8]。

基于模型的迁移学习核心目标就是捕捉源域模型中有助于目标域模型高效学习的部分[9]，可大致分为两类方法：基于共享模型成分的迁移与基于正则化的迁移。基于共享模型成分的迁移主要是通过重新利用源域模型中适用于目标域的部分，或者调节源域模型的超参数来适应目标域任务。基于正则化的迁移主要作用是限制模型的灵活性，避免了模型的过拟合。

3.4. 基于关系的迁移学习方法

与上述三种方法不同，此类方法是通过关系域来实现迁移学习。在关系域中数据并不是独立且均匀分布的，而是由多种关系表示[10]。基于关系的迁移学习目标是建立起目标关系域和源关系域之间关系知识的映射，基于两者之间的关系存在共同规律，可以通过关系特征来传递与域无关的知识。基于关系的迁移学习有两种方法：基于一阶关系的迁移学习和基于二阶关系的迁移。

基于一阶关系的迁移学习就是对于两个相关的关系域，它们可以直接跨域共享一些数据样本之间的关系；基于二阶关系的迁移学习则假设两个相关的关系域，它们之间存在相似性，并且拥有独立于具体关系的通用规则，这些规则可以被抽象的从源域中提取出来，然后迁移到目标域上。

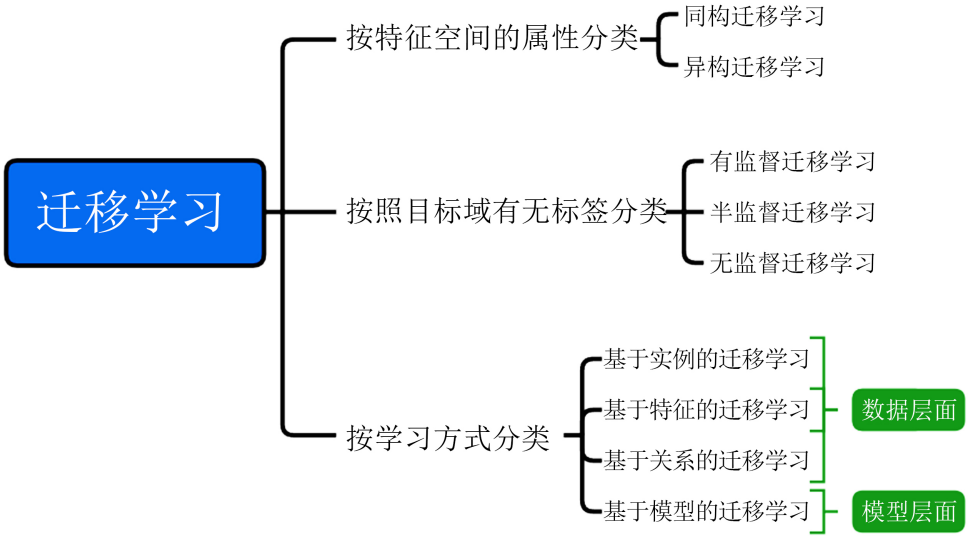

Figure 4. Classification of transfer learning
图 4. 迁移学习的分类

4. 两种基于实例迁移方法的对比

Bagging 的基础训练模型是并行生成的，而 Boosting 则是串行生成的，这也导致两者在权重更新的方式上不相一致。本小结则从训练样本、样本权重、基学习器三个层面对两种方法进行横向对比，结果如表 1 所示。

1) 训练样本

Bagging 从源域中有放回的随机抽取 n 个样本，共抽取 N 轮，形成 N 个源域训练样本子集，对于每个训练样本子集生成一个基础模型；Boosting 则是将所有源域样本整合为一个大的训练集，然后利用该训练集迭代 N 次生成 N 个基础模型。

2) 样本权重

Bagging 在模型训练时并不会对样本权重进行更新；Boosting 则是在每一次迭代中更新整个训练集的样本权重，采用不同方式分别调整源域和目标域样本的权重，使之能更有效的训练模型，提高模型在目

标任务中的表现。

3) 最终模型

Bagging 将并行生成的 N 个基础模型集成为最终模型，其输出是通过对每个基础模型的输出结果进行投票(分类问题)或求取平均值(回归问题)得到的；Boosting 对串行生成的 N 个基础模型依次打分(加权)后集成为最终模型，其输出是每个基础模型输出的加权和。

Table 1. Comparison of instance migration methods based on Bagging and Boosting
表 1. 基于 Bagging 与 Boosting 的实例迁移方法对比

评价层面	Bagging	Boosting
训练样本集	相互独立	与之前迭代结果有关
样本权重	根据与目标域的相似性加权	根据每一次迭代的表现调整
基学习器	并行生成	串行生成

5. 基于迁移学习的钻井机械钻速预测模型设计

5.1. 基于迁移学习的钻井机械钻速预测模型可行性分析

目前传统的机器学习机械钻速预测模型泛化性低，在应用于新井的钻速预测时往往会出现精度显著下降的情况。这是由于传统机器学习模型都是假设数据为独立同分布，而在实钻过程中，不同井会遇到不同的地质情况，因此现场不同井的录井数据并不能满足传统模型的假设。

而上文提到的基于实例的迁移学习方法，恰好能应对我们目前面临的难题。我们每口井的主要录井参数都是一致的，在迁移学习中对应特征空间一致，因此可以把完钻井的录井数据看作迁移学习中的源域，把新钻井看作目标域，利用调整样本权重的思想，结合集成学习，实现模型的迁移。

5.2. 基于迁移学习的钻井机械钻速预测模型的迁移设计

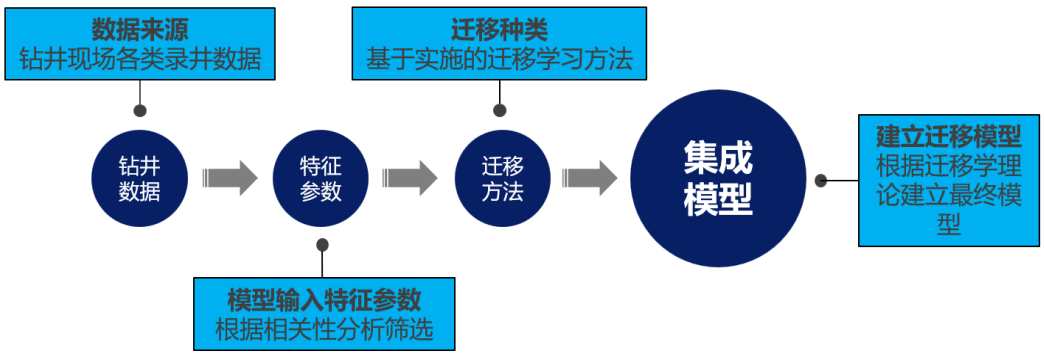

Figure 5. Flowchart of the integrated migration model research idea
图 5. 集成迁移模型研究思路流程图(没有)

针对钻井现场不同井录井数据边缘概率分布不同这一特点，本文模型设计的中心思想是从两个层面进行迁移：集成模型的研究思路见图 5。

- 样本层面利用相关指标衡量源域样本与目标域的相似性，选择相似性高的组成源域训练样本集。
- 结构层面利用筛选出来的源域样本与部分目标域样本训练基学习器，将多个训练好的基学习器在目标域中测试，筛选表现最好的一批作为最终的基础模型，再将基础模型组合成强大的集成模型。

基金项目

重庆市研究生科研创新项目资助；项目编号：CYS23740。

参考文献

[1] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. The MIT Press, Cambridge, 8-49.

[2] 王晋东, 陈益强. 迁移学习导论[M]. 第 2 版. 北京: 电子工业出版社, 2022: 87-92.

[3] Pan, S.J. and Yang, Q. (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[4] Dai, W.Y., Yang, Q., Xue, G.R., *et al*. (2007) Boosting for Transfer Learning. *Proceedings of the* 24*th International Conference on Machine Learning*, San Francisco, 20-24 June 2007, 193-200. https://doi.org/10.1145/1273496.1273521

[5] 李燕. 基于集成学习的多源域实例迁移算法研究[D]: [硕士学位论文]. 昆明: 云南大学, 2018.

[6] Pan, J. (2010) Feature-Based Transfer Learning with Real-World Applications. Ph.D. Thesis, Hong Kong University of Science and Technology.

[7] Zhong, X., Guo, S., Shan, H., *et al*. (2018) Feature-Based Transfer Learning Based on Distribution Similarity. *IEEE Access*, **6**, 35551-35557. https://doi.org/10.1109/ACCESS.2018.2843773

[8] Wang, J., Chen, Y., Feng, W., *et al*. (2020) Transfer Learning with Dynamic Distribution Adaptation. *ACM Transactions on Intelligent Systems and Technology* (*TIST*), **11**, Article No. 6. https://doi.org/10.1145/3360309

[9] Zhang, Y.H., Zhang, Y. and Yang, Q. (2019) Parameter Transfer Unit for Deep Neural Networks. *Advances in Knowledge Discovery and Data Mining*: 23*rd Pacific-Asia Conference*, *PAKDD* 2019, Macau, 14-17 April 2019, 82-95. https://doi.org/10.1007/978-3-030-16145-3_7

[10] Farahani, A., Pourshojae, B., Rasheed, K., *et al*. (2020) A Concise Review of Transfer Learning. 2020 *International Conference on Computational Science and Computational Intelligence* (*CSCI*), Las Vegas, 16-18 December 2020, 344-351. https://doi.org/10.1109/CSCI51800.2020.00065