A METHOD FOR LOGGING DATA RECONSTRUCTION BASED ON TRANSFER LEARNING

by

Yuxuan ZHAO [a], ***Yong GUO*** [a], ***Hao CHEN*** [a], ***Lei YANG*** [b,c*], ***Chuanming XI*** [a], **and** ***Yuqiang XU*** [b,c]

[a] Drilling Engineering Research Institute of Xinjiang Oilfield Engineering Technology Research Institute, Karamay, Xinjiang, China
[b] National Key Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao, China
[c] Shandong Ultra-Deep Drilling Process Control Tech RD Center, Qingdao, China

Original scientific paper
https://doi.org/10.2298/TSCI2404451Z

This paper proposes a logging data reconstruction method based on migration learning, which can reduce the dependence on labeled data and also help to improve the generalization ability of data-driven models. Reconstruction experiments are carried out using data from a block in the Junggar Basin of Xinjiang, and compared with the conventional data-driven long short-term memory network and recurrent neural network methods. The results show that the reconstruction results based on migration learning improve the accuracy by 21%, which is significantly better than the remaining two methods, and proves the feasibility of the method.

Key words: *logging data reconstruction, transfer learning, data-driven*

Introduction

The quality of the acquired logging data directly affects the reliability of interpretation and analysis, affecting exploration and development [1-4]. However, in the construction process, logging data are often missing in well sections due to wall collapse, drilling risks, and cost savings [5]. Meanwhile, for many oil and gas wells that have already been cemented, recapturing the missing log data is extremely challenging and relatively costly [6]. Missing logging data poses a substantial challenge to the development and evaluation of oil and gas reservoirs. Therefore, how to reconstruct the missing log curves has become a new topic to face [7, 8].

Conventional reconstruction methods include empirical formulae (*e.g.*, Garden's formula [9], Faust's formula [10], numerical computation methods such as interpolation [11], *etc.*). However, these conventional reconstruction methods often have limitations. In recent years, machine learning has gradually become an effective means of logging data reconstruction by its powerful data mining capability. Therefore, how to reduce the reliance on labeled data becomes an urgent problem. In addition, although machine learning models perform well on training data, they often suffer from insufficient model generalisation capabilities when facing novel or differently distributed data.

Machine learning creates the possibility for machines to achieve autonomous knowledge learning from data and apply it to solve new problems. Transfer learning, an essential

* Corresponding author, e-mail: Z22020089@s.upc.edu.cn

branch of machine learning, we focus on transferring acquired knowledge to new problems to enhance the ability to solve new problems and increase the speed of solutions. Therefore, the research goal of this paper is to explore the feasibility of applying transfer learning to log data reconstruction.

Transfer component analysis

The transfer component analysis (TCA) transfer learning algorithm tries to make it possible to reduce the data distribution differences significantly while preserving the data attributes by mapping the source and target domain data into the reproducing kernel Hilbert space (RKHS). The TCA is a transfer learning method based on maximum mean discrepancy (MMD), whose MMD distance can be represented:

$$\mathrm{DIST}\left(X_s, X_t\right)=\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\Phi\left(x_{si}\right)-\frac{1}{n_t}\sum_{j=1}^{n_t}\Phi\left(x_{tj}\right)_H\right\| \tag{1}$$

where n_s and n_t are the number of samples in the source and target domains, respectively, and $\|*\|_H$ is the RKHS. However, eq. (1) cannot be computed directly, and some kernel method is required to transform the samples into the mapping space. To map the data in the source and target domains into a shared low dimensional potential space, TCA introduces the kernel matrix K and the distributional difference matrix L_{ij}:

$$K=\begin{bmatrix} K_{ss} & K_{st} \\ K_{ts} & K_{tt} \end{bmatrix} \tag{2}$$

and

$$L_{ij}=\begin{cases} \dfrac{1}{n_s^2}, & x_i, x_j \in X_s \\ \dfrac{1}{n_t^2}, & x_i, x_j \in X_t \\ -\dfrac{1}{n_s n_t}, & \text{otherwise} \end{cases} \tag{3}$$

where K_{ss} and K_{tt} are the kernel matrices on the source and targetdomains. The K_{ss} and K_{tt} are the kernel matrices across domains.

The distance in eq. (1) can be as $tr(K) - \lambda tr(K)$, where the first of these minimizes the distance between the distributions, the second maximizes the variance in the feature space, and $\lambda(\lambda \geq 0)$ is the weight parameter:

$$\begin{gathered} \min_W tr\left(W^T KLKW\right)+\mu tr\left(W^T W\right) \\ s.t.\ W^T KHKW = I_m \end{gathered} \tag{4}$$

where $\mu > 0$ is the weight parameter, I_m – the unit matrix of size $m\times m$, and H – the centre matrix defined as

$$H = I_{n_s+n_t} - \left(\frac{1}{n_s+n_t}\right)11^T$$

The data are downscaled to the new source and target domain data.

Long short term memory neural networks

Long short-term memory network (LSTM) is an improved algorithm for recurrent neural network (RNN). It combines short-term memory with long-term memory by introducing a gating mechanism, which solves the limitation that traditional RNN can only process short-term input information, and at the same time solves the problems of gradient vanishing and gradient explosion a certain extent. Figure 1 illustrates the model structure of LSTM.

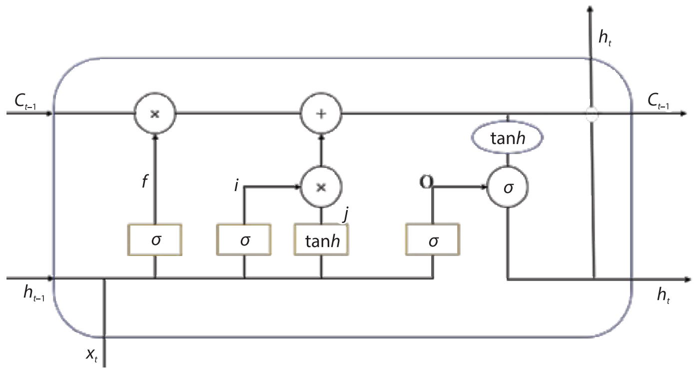

Figure 1. The LSTM model structure

Here, *Ct* – 1 and *Ct* are the hidden node state, *ht* – 1 and *ht* denote the previous and current sequence hidden node outputs, respectively, *xt* is the current sequence hidden node input; *f*, *i*, *j*, *o* are the process quantities, and σ is the sigmoid non-linear activation function, and tanh represents the hyperbolic tangent function.

The TCA-LSTM model

In this paper, the transfer learning algorithm TCA is combined with the LSTM neural network to form the TCA-LSTM model, and fig. 2 shows the specific structure of the TCA-LSTM model in detail.

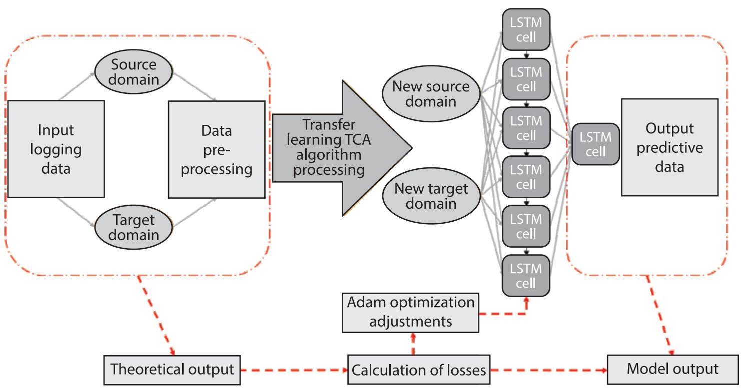

Figure 2. The TCA-LSTM model structure

Model evaluation indicators

The model evaluation indexes chosen in this paper are mean absolute error (MAE) and root mean squared error (RMSE). The formulas for MAE and RMSE are given:

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

and

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{true}} - y_{\text{pred}}\right)^2} \tag{6}$$

where y_{true} is the true sample value, y_{pred} – the predicted sample value, and n – the sample size.

Logging data reconstruction

In this paper, the logging data from six wells in the region were collated and data pre-processed, and it was found that there were outliers and some missing segments in the acoustic time difference data in some of the logging data, which caused a serious obstacle to the subsequent formation pressure prediction and post-drilling risk evaluation of drilling projects. Therefore, six wells from the same block were used for training tests, with well M2 as the target domain and the remaining wells as the source domain.

To test the performance of the TCA-LSTM model, the following two sets of experiments were conducted:
– A feature log in a well is completely missing, the missing log is reconstructed by using the complete log of the well and the complete log of the neighboring well.
– A feature in a well is partially missing in a section, the complete log of the well and the complete log of the neighboring well are used to reconstruct the missing log.

According to the experimental model evaluation metrics MAE and RMSE are compared with conventio*nal LSTM and CNN neural network methods.*

Experiment 1

By analysing the logging data in the west basin depression area, this paper artificially deletes the acoustic time-difference data in well M2, and uses the logging data collected from

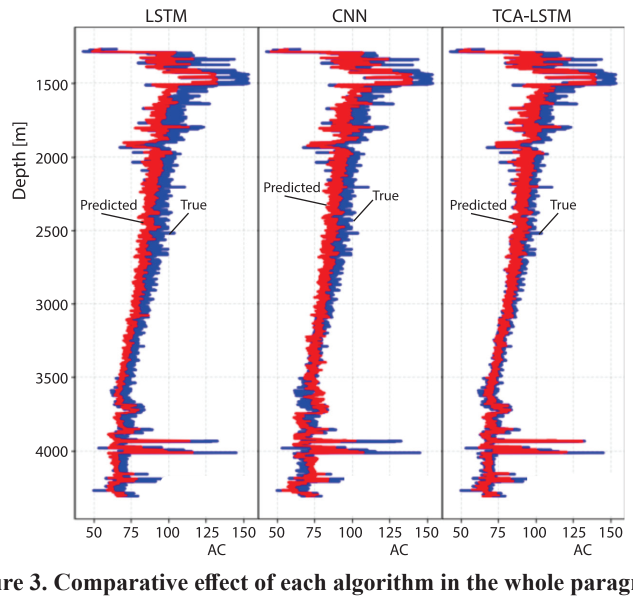

Figure 3. Comparative effect of each algorithm in the whole paragraph

the remaining six wells as the source domain for training, and performs data reconstruction of the acoustic time-difference data from well M2. Figure 3 shows the effect of various methods, and fig. 4 shows the model metrics of each method.

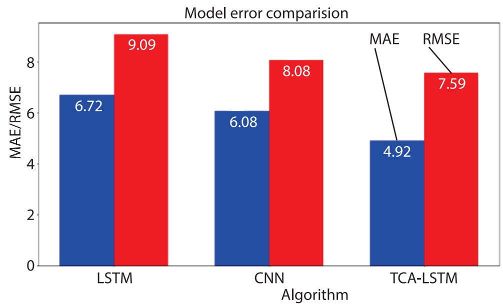

Figure 4. Comparison of model error of each algorithm in the whole segment

Experiment 2

The logging data collected from the remaining five wells were used as the source domain for model training, the missing segments were altered, and the acoustic time-difference data from well M2 was data reconstructed by artificially deleting the acoustic time-difference data from the 2000-3000 segments in well 2. Figure 5 demonstrates the effect of various methods, and fig. 6 shows the model metrics of each method.

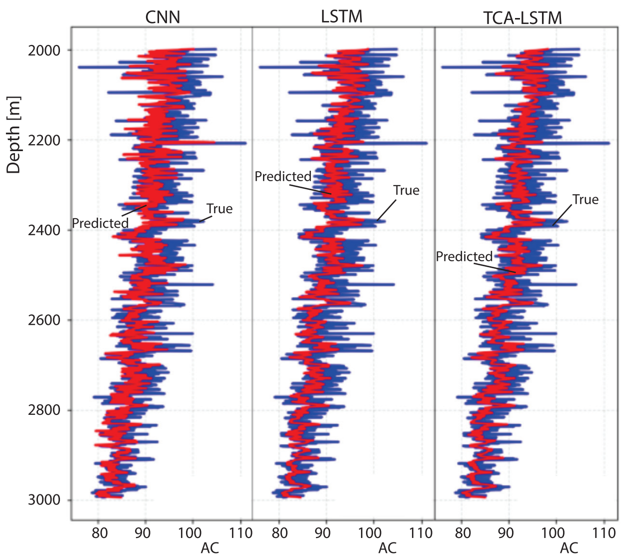

Figure 5. Comparative effectiveness of algorithms in missing segments

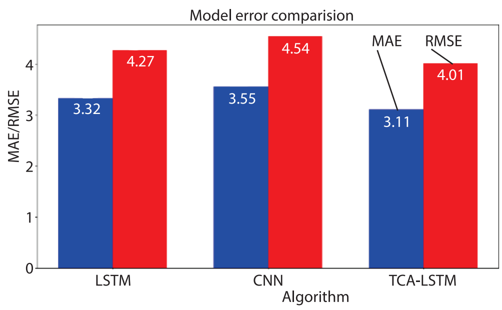

Figure 6. Comparative effectiveness of algorithms in missing segments

Conclusion

Through detailed investigation and research, we found that when there is some difference between source domain data and target domain data, it will lead to poor performance of traditional models on target domain data. To solve this problem, we introduce a transfer learning method, which can improve the performance of the model on the target domain data with the knowledge of the source domain data. In this context, we adopt the TCA transfer learning algorithm. The experimental results show that the TCA transfer learning algorithm performs well in the field of logging data analysis, and can reconstruct logging data more accurately, thus providing more accurate logging data. This ability to accurately reconstruct logging data provides efficient operational guidance at the construction site, as well as a more reliable basis for oil and gas reserve assessment, providing strong support for exploration and development decisions.

Acknowledgment

The authors would like to acknowledge the academic and technical support of China University of Petroleum (East China). This paper is supported by the National Natural Science Foundation of China (52274024, 52074326) and also supported by the Shandong Provincial Natural Science Foundation (ZR2023YQ045).

Reference

[1] Yong, S., *et al*., *Logging Data Processing and Comprehensive Interpretation*, China University of Petroleum Press, Dongying, China, 2007
[2] Zhang, L., *et al*., Well Data-Driven Thin Interlayer Reservoir Prediction Technique in the Bohai a Oilfield, *Xinjiang Oil and Gas*, *17* (2021), 2, pp. 7-12
[3] Yang, W., *et al*., Oil Logging Reservoir Recognition Based on TCN and SA-BiLSTM Deep Learning Method, *Engineering Applications of Artificial Intelligence*, *121* (2023), 1, 105950
[4] Liu, J.-J., *et al.*, Integrating Deep Learning and Logging Data Analytics for Lithofacies Classification and 3-D Modelling of Tight Sandstone Reservoirs, *Geoscience Frontiers*, *13* (2022), 1, ID101311
[5] Xie, X., *et al*., A Systematic Study of Borehole Quality Evaluation, *Xinjiang Oil and Gas*, *16* (2020), 2, pp. 33-37
[6] Wang, J., *et al*., Logging Curve Prediction Based on a CNN-GRU Neural Network, *Geophysical Prospecting for Petroleum*, *61* (2022), 2, pp. 276-285
[7] He, X., *et al.*, Log Facies Recognition Based on Convolutional Neural Network, *Oil Geophysical Prospecting*, *54* (2019), 5, pp. 1159-1165
[8] Kuang, L., *et al*., Application and Development Trend of Artificial Intelligence in Petroleum Exploration and Development, *Petroleum Exploration and Development*, *48* (2021), 1, pp. 1-11
[9] Huang, X., *et al*., An Exploration of the Conditions for the Application of Garden's Formula, *Xinjiang Petroleum Geology*, *3* (2000), 3, pp. 233-234
[10] Chen, G., *et al*., Application of Faust Equation in Rebuilding of Acoustic Curve, *Progress in Exploration Geophysics*, *26* (2005), 2, pp. 125-128
[11] Kong, Y., *et al.*, Application of Interpolating Based on Excel to Logging Data Processing, *Petroleum Instruments*, *21* (2008), 2, pp. 58-60

Paper submitted: March 24, 2024
Paper revised: March 29, 2024
Paper accepted: May 11, 2024