Data generation and verification of pipeline transportation based on integration of simulation and DBSCAN algorithm

Xinru Zhang [1, 2, a], Lei Hou [1, 2, b], and Zuoliang Zhu [1, 2, c]

[1]School of machine and transportation, China University of Petroleum in Beijing, China;

[2] National Engineering Laboratory for Pipeline Safety/MOE Key Laboratory of Petroleum Engineering, China University of Petroleum in Beijing.

[a] zhangxr_cup@163.com, [b] houleicupbj@126.com, [c] zhuzuoliang1995@gmail.com

Abstract. Through the analysis and mining of historical data, machine learning method can be used to obtain high accuracy prediction effect without establishing a complex physical model in equipment fault diagnosis, operation condition prediction, and pipeline energy consumption analysis. In the oil and gas pipeline system, the machine learning model unable to gain an ideal training effect with the data set, because of confidentiality of data, imperfect data acquisition technology, low frequency of abnormal working conditions, and other factors. In this paper, aiming at the operation energy consumption of a crude oil pipeline, the power consumption of oil pump unit is simulated by software, which can expand the data. The quality of simulation samples has a great effect on the training results. A DBSCAN algorithm based on Mahalanobis distance is proposed to evaluate the reliability of simulation samples and identify abnormal simulation samples, given the characteristics of virtual samples in pipeline transmission simulation, such as no real value control, feature correlation, and high dimension. Examples have shown that the fitting ability of the model can be improved after the simulation samples for eliminating abnormal data are added to the training set, which provides a new method for the generation and verification of simulation samples.

Keywords: machine learning; crude oil pipeline; energy consumption prediction; simulation sample; DBSCAN algorithm.

1. Introduction

Machine learning can be used for the analysis and mining of historical data, discovering correlations between data, overcoming the difficulty of completely relying on theoretical knowledge modeling, and is an effective means to solve complex engineering problems. It has broad application prospects in equipment fault diagnosis [1-3], energy consumption prediction [4-5], operating condition prediction [6-7], and other aspects. Machine learning models are very sensitive to the quantity and quality of training samples, and a small number of samples cannot achieve good training results. However, in some application scenarios, the data available is usually limited. For example, pipeline systems that have been in service for a long time are limited by the technology used during construction, resulting in low data collection levels or low frequency of operating conditions such as pipeline leaks and equipment failures, lacking appropriate historical data. To increase the effective number of training samples, scholars have proposed data generation techniques to improve model prediction performance by expanding training sample.

Modeling and simulating in simulation software based on existing parameters and data can generate virtual samples of assumed operating conditions, and the obtained virtual samples have physical significance. The quality of simulation data has a direct impact on the performance of the model. Low consistency between simulation data and real data can lead to a decrease in the predictive performance of the model. It is necessary to evaluate the quality of simulation data. The generated simulation samples are simulation data for some assumed operating conditions, without corresponding real values of the operating conditions, and traditional prediction and evaluation indicators such as MSE (Mean Square Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error) cannot be used. For such samples, distance discrimination is used to verify whether the two samples are similar through the distance between the simulated sample and the real sample. Commonly used distances include Euclidean distance, standard Euclidean distance,

Markov distance, etc. [8]. Simply using distance for discrimination requires a threshold of distance, which needs to be determined by the operator based on experience and has a certain degree of subjectivity. Distance based clustering algorithms can allocate a given set of samples to different sets based on similarity, resulting in high similarity among samples within the same set [9]. If the clustering method categorizes simulated samples and real samples into the same set, it indicates that the simulated samples are similar to the real samples, thereby verifying the reliability of the simulated samples.

In order to solve the problem of few application scenario data samples, this paper uses TLNET software to simulate the power consumption of pump units in a heat pump station of an oil pipeline based on existing data, generate simulation samples, identify abnormal data based on Mahalanobis distance and DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm, verify the reliability of generated simulation samples, and add the simulation samples excluding abnormal data to the data set, Used to train BP neural networks and evaluate the impact of adding simulation samples on model prediction performance.

2. Generation of simulation samples

2.1 Simulation software

Based on the existing data of a certain crude oil long-distance pipeline, the power of the pump unit in the heat pump station is simulated using TLNET software. TLNET is a component of the Pipeline Studio series pipeline simulation software produced by ESI (Energy Solution International) in the UK that performs liquid pipeline simulation and simulation. TLNET's liquid pipeline simulation engine has been validated by the industry and has an intuitive graphical environment, providing various types of equipment for modeling [10]. It has been widely used in operation scheme selection [11], leakage simulation [12], optimization control [13], and other aspects.

2.2 Pipeline data

A certain pipeline is an insulated crude oil pipeline in China, with a total length of 361.2 km and a design capacity of 900~1000 × 10^4 t/a. Simulate one of the heat pump stations and inter station pipelines, with the specifications of the inter station pipelines is φ508×7.1 mm, made of L450M spiral welded steel pipe.

The data from July 2019 to April 2020 between the stations is selected, and after removing the missing samples, a total of 300 sets of real samples are used as the basis for establishing the simulation model and verifying the simulation samples afterwards. Some of the real samples are shown in Table 1.

Table 1 Partial real samples

Date	Throughput/t	Average inlet pressure /MPa	Average outlet pressure /MPa	Next station inlet pressure /MPa	Average outlet temperature/ ℃	Next station inlet temperature /℃	Ground temperature /℃	Total power consumption /10^3kW·h^{-1}
2019/07/01	17906	1.95	4.58	1.51	38.8	36.20	18.1	24380
2019/08/01	23925	0.75	5.85	1.01	40.6	38.81	20.9	55586
2019/09/01	23606	1.54	6.71	1.46	40.2	38.49	21.8	55045
2019/10/01	20170	0.37	5.71	1.18	38.8	37.25	21.3	50672
2019/11/01	19071	2.17	4.85	1.61	39.6	36.36	19	24864

2019/12/01	21609	0.52	5.77	0.61	40.8	36.87	15.1	52753
2020/01/01	19886	2.27	4.94	1.35	38.1	36.43	10.9	26585
2020/02/01	18543	1.24	3.89	0.73	39.2	34.43	7.8	24382
2020/03/01	31296	1.24	8.02	0.75	43.8	41.8	6.6	96203
2020/04/01	20032	1.65	4.33	1.38	38.8	38.1	7.6	28241

2.3 Establishment of simulation models

Establish a simulation model based on existing pipeline information and oil product data. Use Tabular Fluids to simulate pipeline oil products, set Class as Crude in Tabular Fluids settings, and input viscosity temperature curve, Bulk modulus curve, heat capacity curve, and thermal expansion coefficient curve of oil products; Simulate the pipeline between stations using Pipe, input parameters such as length, diameter, and wall thickness in the Pipe settings, use Survey to import the elevation data of the pipeline, calculate the range of pipe wall roughness and heat release coefficient based on the recommended values in the specifications, and adjust it based on on-site data. Determine that the pipe wall roughness is 0.0654 mm, and the heat release coefficient is 2.12 W/(m2·K); Use Centrifugal Pump to simulate the pump unit of a heat pump station, and import the head characteristic curve and efficiency characteristic curve of the centrifugal pump unit through Centrifugal CPID; Use Supply and Delivery to simulate the outbound and inbound flow of crude oil. In order to improve the calculation speed of the model, the heating furnace was ignored during modeling, and instead, the inlet temperature and ground temperature were directly set for thermal calculation, without affecting the calculation accuracy of the power consumption of the pump unit.

The specific model is shown in Fig. 1, where the boundary condition of the inlet EIN is Maximum Pressure, the value is the average sink pressure of the heat pump station, and the Fluid Temperature is set as the average outlet temperature of the heat pump station; The control mode of the pump station is set to Max Flow, and the value is the pipeline throughput; The boundary condition for the exit EOUT is Min Pressure, and the value is the next station inlet pressure.

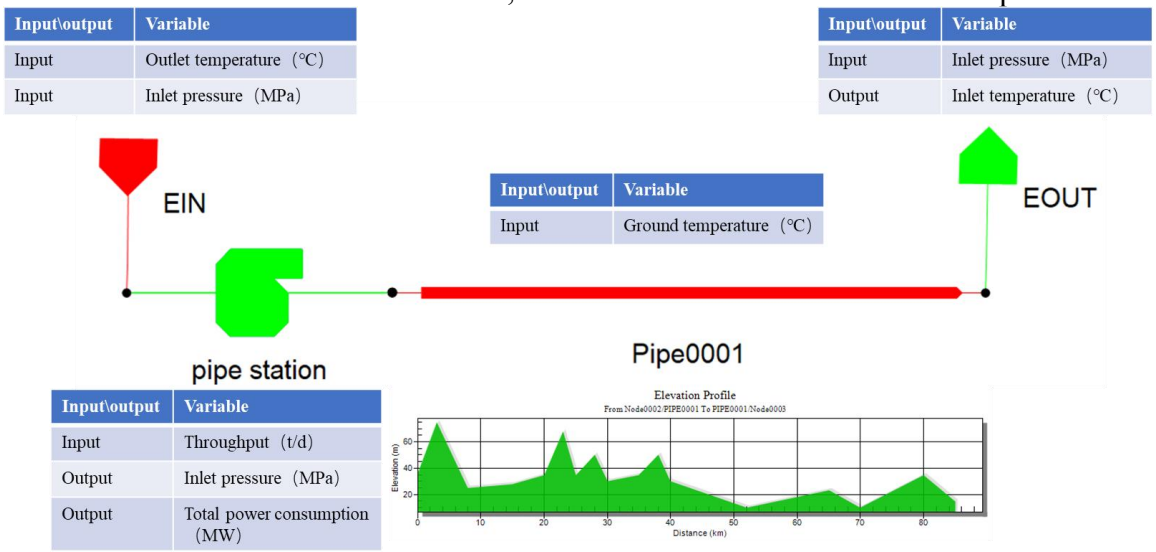

Fig. 1 Simulation model

2.4 Generation of simulation data

Using the constructed simulation model, input variables are determined based on the variation range of on-site data such as throughput, average sink pressure, and next station inlet pressure,

generating simulation data for different simulation conditions. To explore the impact of the number of simulation samples in the dataset on neural network training, 120, 240, 360, and 480 sets of simulation data were generated based on different input variables. The specific input variables are shown in Table 2.

Table 2 Input variables

Number of simulation samples	Throughput/t	Average inlet pressure /MPa	Next station inlet pressure /MPa	Average outlet temperature /°C	Ground temperature /°C
120	800,1000,1200,1400,1500	0.3,1,1.7,2.4	0.5,1.5,2.5	40	7,20
240	750,800,900,1000,1100,1200, 1300,1400,1500,1550	0.3,1,1.7,2.4	0.5,1.5,2.5	40	7,20
360	750,800,900,1000,1100,1200, 1300,1400,1500,1550	0.3,1,1.7,2.4	0.5,1.5,2.5	40	7,14,20
480	750,800,900,1000,1100,1200, 1300,1400,1500,1550	0.3,1,1.7,2.4	0.5,1,1.5,2.5	40	7,14,20

3. Verification of simulation samples

3.1 Verification method

Due to the fact that simulation samples simulate hypothetical operating conditions, the output of the simulation does not correspond to the true values of the operating conditions as a reference, and is a high-dimensional sample containing multiple variables, traditional predictive evaluation indicators are not suitable for validation of simulation samples. This article uses Markov distance to quantitatively characterize the similarity between samples. A density clustering algorithm based on distance is used to cluster simulated and real samples, excavate the inherent correlation between samples, eliminate simulation samples with poor similarity to real samples, and evaluate the data quality of simulation samples.

3.2 Distance Calculation

Whether the distance between samples can accurately and comprehensively represent the relationship between samples directly affects the accuracy of clustering, and distance calculation generally uses Euclidean distance and Mahalanobis distance. Compared to Euclidean distance, Mahalanobis distance can consider the connections between various features within the sample and is scale independent [14]. Due to the interconnection of features such as outbound pressure and inbound pressure between the simulation samples and the real samples, the data of features such as throughput and temperature differ by two to three orders of magnitude. Therefore, the distance between the samples is represented by Mahalanobis distance, and the calculation formula for Mahalanobis distance is as follows:

$$dis = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)} \tag{1}$$

Where X is the sample; S is the population sample Covariance matrix; i, j=1, 2,..., n (n is the number of sample features).

3.3 Sample similarity verification based on DBSCAN algorithm

The distance between samples can represent the differences between different samples. In the face of a large number of simulated and real samples, judging similarity based on a single threshold cannot reflect the correlation between samples, and the setting of the threshold also carries some subjectivity. The clustering method assigns similar data points to the same cluster based on the distance between samples, and verifies the similarity between simulated and real samples by

determining whether they can be assigned to the same cluster. Clustering methods include prototype clustering, density clustering, and Hierarchical clustering [15]. As a typical density clustering algorithm, DBSCAN algorithm, compared with other clustering algorithms, has the advantages of not having to specify the number of clusters in advance, being suitable for dense non convex data sets, and being able to find noise points during clustering. For the validation of simulation samples, due to the inability to estimate the number of clusters, the small distance between samples, and the need to eliminate simulation samples that differ greatly from the real samples, the DBSCAN algorithm is suitable for the validation of simulation samples. Some literature [16] points out that in the face of high-density data, the DBSCAN algorithm has certain advantages in the accuracy and efficiency of cluster partitioning.

Mix 300 sets of real samples with simulation samples to form 420, 540, 660, and 780 sets of samples. Use the DBSCAN model with determined parameters to classify the sample set. The classification results are shown in Table 3. The vast majority of simulation samples and real samples are divided into the same cluster, and only a small number of heterogeneous virtual samples need to be removed. The number of simulation samples after removal is 412, 535, 655, and 775, The similarity between simulated samples and real samples is relatively high.

Table 3 DBSCAN classification results

Number of total samples	Number of simulation samples	Number of simulation samples in different clusters	Number of training set samples
420	120	8	412
540	240	5	535
660	360	5	655
780	480	5	775

4. The Impact of Simulation Samples on Machine Learning Models

Fig. 2 shows the trend of changes in the prediction index MAPE when simulation data is added to the training set. The specific prediction and evaluation indicators of the model using different training sets are shown in Table 5. The analysis results show that after adding simulation data, all prediction indicators have improved compared to those without simulation data, but the prediction indicators of adding 775 groups have decreased compared to adding 655 groups of simulation data, This is because more simulation data may mean simulating more extreme operating conditions (such as low ground temperature and low flow), resulting in a decrease in the model's ability to fit conventional operating conditions.

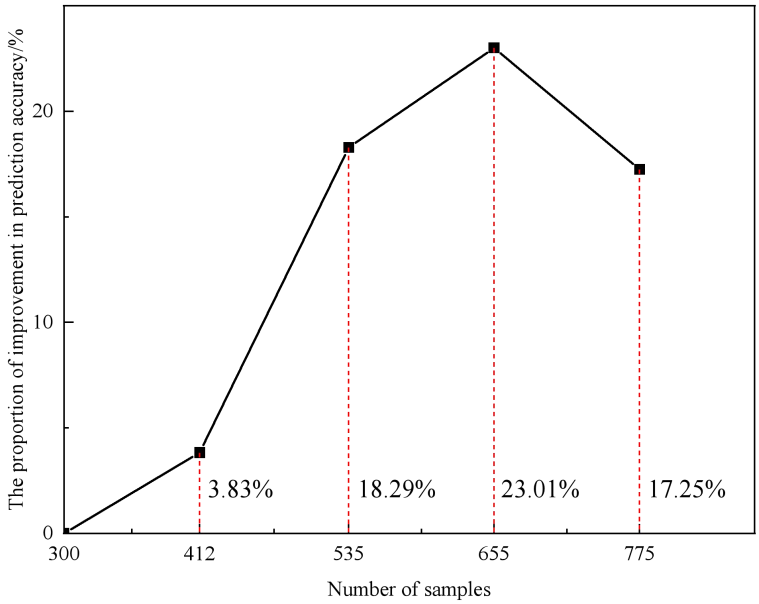

Fig. 2 The Impact of Simulation Samples on Machine Learning Models

5. Summary

(1) To solve the problem of insufficient pipeline data samples leading to poor training performance of machine learning models, based on existing pipeline data, TLNET software is used to simulate the power consumption of a pumping station in a long-distance pipeline, generate simulation samples, and expand the data. A DBSCAN algorithm based on Markov distance is proposed to address the characteristics of no real value comparison, feature correlation, and high dimensionality of simulated samples. This algorithm can identify abnormal simulated samples and verify the similarity between simulated and real samples.

(2) Add simulation samples that eliminate outliers to the dataset for training the BP neural network. The results show that after adding 112, 235, 355, and 475 sets of simulation data to the dataset, the MAPE decreases by 0.22%, 1.05%, 1.32%, and 0.99%, respectively. This proves that adding simulation samples to the dataset can effectively improve the model's prediction ability and reduce prediction errors, Overfitting the "overfitting" phenomenon of the prediction model when the initial sample is small.

(3) For certain application scenarios in the oil and gas transportation industry, such as high data confidentiality, incomplete data collection technology, and low frequency of abnormal operating conditions, based on the collected data, software is used to simulate the application scenarios, generate simulation samples, and remove unreasonable simulation samples through clustering method. This article takes the energy consumption of crude oil pipelines as an example to verify that this method can expand the dataset and improve the prediction performance of machine learning models.

References

[1] LI Zixi. Study on fault diagnosis of pipeline leakage signal based on time frequency [D]. Northeast Petroleum University, 2016.

[2] YU Deliang, LI Yanmei, DING Bao, et al. The MMAGA-RBF fault diagnosis method for submersible plunger pump [J]. Journal of Harbin Institute of Technology, 2017,49 (09): 159-165

[3] Maamar Ali Saud ALTobi,Geraint Bevan,Peter Wallace,David Harrison,K.P. Ramachandran. Fault diagnosis of a centrifugal pump using MLP-GABP and SVM with CWT[J]. Engineering Science and Technology, an International Journal,2019,22(3).

[4] SHI Yao. Research on energy consumption analysis and optimization of joint station based on big data mining [D]. China University of Petroleum (East China), 2018

[5] XU Lei, HOU Lei ,LI Yu, et al. Research into prediction of energy consumption of crude oil pipelines based on machine learning [J]. Petroleum Science Bulletin,2020,03:576-586

[6] ZHANG Xu. Study on classification of gas transmission pipeline based on VMD and neural network [D]. Northeast Petroleum University, 2018

[7] GUAN Fusheng. Research on identification technology of oil pipeline condition based on GIF Elman neural network [D]. Northeastern University, 2011

[8] WU Yi, LI Yongle, HU Qingjun. Applied mathematical statistics [M]. National University of Defense Science and Technology Press, 1995.

[9] TANG Xinyao, ZHANG Zhengjun, CHU Jie, et al. Density peaks clustering algorithm based on natural nearest neighbor [J/OL].Computer science:1-12[2020-11-29].http://kns.cnki.net/kcms/detail/50.1075.TP.20201104. 1623.004.html.

[10] Energy Solutions International.Pipeline Studio (Version 4.0) .2015.

[11] Li Xin Wei,Lu Ying Zhang,Yu Wang. Simulation of Operation Scheme of Su Cuo Buried Oil Pipeline[J]. Advanced Materials Research,2012,1479.

[12] TONG Shujiao, WU Zongzhi, WANG rujun, et al. Simulation and analysis of transient leakage for long distance oil pipelines based on TLNET [J]. Safety and Environmental Engineering, 2016,23 (01): 128-132 + 139

[13] ZHANG Lei, ZHANG Chenyuan. Optimal selection of pump start-up combinations in pumping stations [J]. Chemical Engineering & Equipment, 2016, (2): 111-113

[14] WU Qing, ZHANG Yu, ZANG Boyan, et al. Possibilistic entropy clustering algorithm based on Mahalanobis distance [J]. Computer Simulation, 2019,36 (12): 240-243 + 312

[15] ZHOU Zhihua. Machine learning [M]. Tsinghua University Press,2016.

[16] FANG Zonghua, WANG Wenfeng, DONG Jianhua, et al. Research on DBSCAN algorithm based on Hermitian interpolation [J]. Journal of Nanchang Institute of Technology, 2020,39 (04): 80-84