A Deep Neural Network Based Feature Learning Method for Well Log Interpretation

Liyuan Bao[1], Xianjun Cao[2], Changjiang Yu[2], Guanwen Zhang[1(✉)], and Wei Zhou[1]

[1] Northwestern Polytechnical University, Xi'an 710072, China
guanwen.zh@nwpu.edu.cn
[2] China Petroleum Logging Co., Ltd., Xi'an 710077, China

Abstract. Well log interpretation is an important task in the process of petroleum logging. It is able to help the researchers to determine the residual oil volume and to improve the petroleum productivity efficiency. Well log interpretation requires the synthesis of a large amount of data, and it is difficult to manually browse the data from a global perspective. It is urgent to introduce big data analysis methods to deal with the complex oil well logs data. The accuracy of logging interpretation greatly depends on the logging features selection and representation. However, the conventional methods using expert experiences easily lead to feature incomplete problem and affects the interpretation results. In this paper, we propose a deep neural network based feature learning method for well log interpretation. Firstly, we select original features of the well log data according to the physical characteristics of well logging sensors. And then, we formulate a deep neural network based autoencoder model to explore the intrinsic representation of original features. At last, we utilize linear SVM classifier on well log interpretation problem to evaluate the proposed feature learning method. The experimental results demonstrate that the classification accuracy by using learned feature representation increase to 99.8% compared with that of 74.6% by using original feature representation.

Keywords: Well logging interpretation · Deep neural network · Feature learning · Autoencoder

This work was supported in part by the National Key R&D Program of China (No. 2018AAA0102801 and No. 2018AAA0102803), in part of the National Natural Science Foundation of China (No. 61772424, No. 61702418, and No. 61602383), and in part by 13th Five-Year Plan of China National Petroleum Corporation Limited (2019A-3610).

B. Li et al. (Eds.): IoTaaS 2020, LNICST 346, pp. 543–556, 2021.
https://doi.org/10.1007/978-3-030-67514-1_43

1 Introduction

Nowadays, oil has become a strategic material that all countries attach great importance to. However, petroleum energy is facing some problems such as uneven distribution and difficult mining technology. Logging technology is an indispensable method for accurately discovering oil and gas layers and finely describing oil and gas reservoirs, and it is able to help to achieve more efficient production. Logging interpretation requires the synthesis of a large amount of data, and it is difficult to manually browse the data from a global perspective. It is urgent to introduce big data analysis technology to deal with complex oil well logs data. By introducing the artificial intelligence methods into the logging interpretation, the establishment of intelligent logging interpretation will improve the automation and accuracy of logging interpretation. The classification effect of deep learning and machine learning depends on the expressive ability and separability of logging features. Therefore, learning feature representation of logging interpretation is very an important research field.

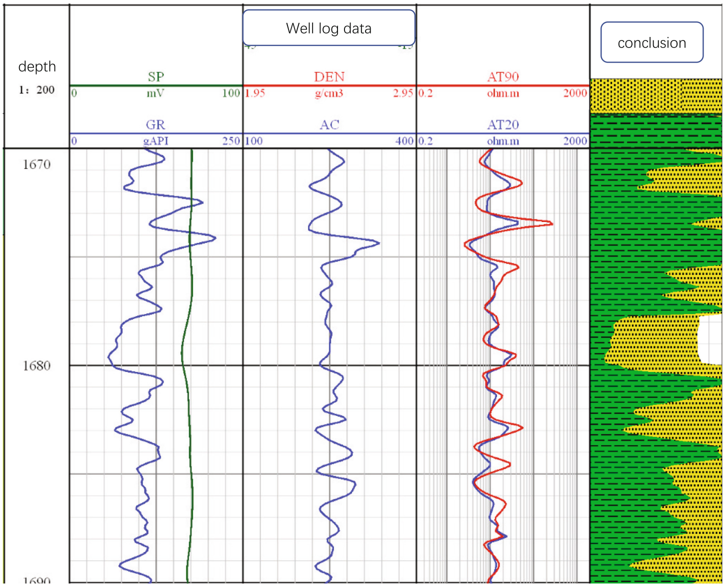

Fig. 1. An example for well logging interpretation application by using manual designed feature representation.

The logging feature representation consists of manual designed feature method and learning feature method. Experts get logging interpretation by drawing intersection diagrams of well logs data, as the example shown in Fig. 1. The well logs data is generally used to determine the main features in the logging. However it greatly requires the expert knowledge and experience, and it easily leads to insufficient use of logging features and inaccurate logging interpretation results. In [22], Liu et al. tested 127 mudstone samples in Shanxi Formation by X-ray diffraction (XRD), scanning electron microscope (SEM), and gas content. They used the logging curve superposition method and reservoir parameter calculation equations to qualitatively identify and quantitatively evaluate gas-bearing mudstone reservoirs in 4 gas wells, and achieved good results. However, this method has the problem of time-consuming data collection and analysis, and the analysis results have large human errors.

In this paper we proposed a deep neural network based feature learning method for well log interpretation. On the basis of deep neural network, we formulate a autoencoder model to explore the intrinsic features of well log data. The encoded feature representation is learned by minimizing the reconstruction error between the original log data and predicted data. The proposed method provides a high-dimensional representation method for well logs data features. The learned feature representation is expected to contribute the linear separability for classification problem in well log interpretation problem. The method proposed consists of three parts. Firstly, we select the original features of the well logs data according to physical characteristics of the logging sensors. And then we construct the encoder network to perform feature learning on the basis of original features. Finally, we utilize linear SVM as a classifier to evaluate the proposed feature learning method. The experimental results demonstrate that the classification accuracy by using learned feature representation has been greatly improved from 74.6% to 99.8% compared with that by using the original well data.

2 Related Work

Logging interpretation produces huge production benefits in the field of oil and gas reservoir exploration. It has become an indispensable part of modern oil and gas exploration [2,7]. The widely used methods in well logging interpretation consists of artificial neural network methods, machine learning methods, statistical analysis methods, etc. [10,13,14]. The latest mathematical analysis knowledge has also been applied to well logging interpretation. These emerging information processing technologies have greatly promoted the vigorous development of the logging interpretation industry [11]. At present, there are two main group methods for presenting features of well logs data, i.e., the manual designed feature method and feature learning methods.

In the group of feature evaluation and selection in well logging interpretation, researchers proposed various method to analyze sensitive factors and to deal with the complex features in well logs data [5,12]. Li et al. used a Relief-F feature selection for processing a large number of high-dimensional data sets [6]. They tried to select sensitive attributes for multiple attributes extracted from logging and seismic data. By comparing the prediction results, they showed the reliability and accuracy of the sensitive factors for the Relief-F feature selection. Principal component analysis (PCA) [21] is very widely used in well logs data for dimensionality reduction and sensitive attribute selection. By mapping highly redundant information to low-dimensional principal component space, PCA simplifies the data model while retaining effective information as well as reducing staff judgment difficulty. Linear discriminant analysis [4] is a supervised learning method. It pays more attention to the distance between different classes, so it is more common to be applied in the dimensionality reduction scenario of the gap between the classifications.

In the group of feature learning method, the autoencoder is also a common method for data feature extraction. Seeun Jo [3] and others used the autoencoder to extract the features of the sample in the spectrum to improve the recognition accuracy. In [1], a new stacked local preserving autoencoder is proposed, which can better preserve the local data structure. The conventional machine learning methods show excellent classification performance on many classification problems [15]. In [16], Wang et al. proposed the application of stacked sparse autoencoders based on PCA and SVM for power system line trip fault diagnosis. A sentiment analysis method based on kmeans and online migration learning was proposed in [8], which achieved good results in error rate and classification accuracy. As an emerging information processing technology, deep neural network has been widely used in many engineering fields. The application of deep neural network in the field of logging interpretation has also achieved convincing performance. It introduce the contribution to the solution of many difficult problems in the field of logging interpretation [9]. Zhang et al. used artificial neural network (ANN) to establish the non-linear correspondence between reservoir lithology, physical properties, and oil-bearing parameters [17]. Their method is able to predict the lithology, physical properties, and oil-bearing parameters of the reservoir. The comparison between the network prediction results and other oil testing and analytical laboratory results shows that the ANN method is feasible and effective. It makes a great reference value for oilfield exploration and development.

Manual designed features have problems that are not conducive to promotion and generalization. The learning feature has the problem of low logging interpretation accuracy. In response to the above problems, this article proposes a method of learning features based on autoencoder. This method can learn the characteristics of the data from the logging data and solve the problem of generalization and promotion of the characteristics. We selected data from a well in Changqing Oilfield to verify the effectiveness of our proposed method. The classification accuracy and clustering effect of this method are greatly improved.

3 The Proposed Method

In the following sections, we first introduce the original features of the well log data with respective to physical characteristics of the logging sensors. And then we formulate a deep neural network based autoencoder model to learn the feature representation for the original features. At last, we demonstrate the application by using the learned features for logging interpretation. The illustration of our proposed method is shown in Fig. 2.

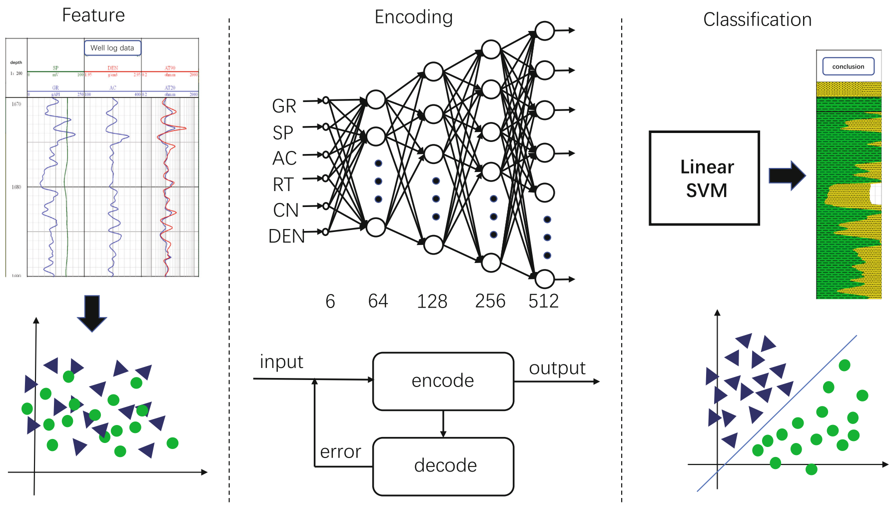

Fig. 2. The illustration of our proposed method. We first select original features for logging data according to the physical characteristics of the sensors. These features are generally linear non-separable. The original features are then used for further feature learning with autoencoder model on the basis of deep neural network. By the minimizing the reconstruction error between original features and predicted features, the learned feature representation is expected to persevere intrinsic distinctiveness of the well log data. The resulting feature representation can therefore contribute the classification for the well log interpretation application.

3.1 Original Feature Selection

There are many original features can be used for well log data representation according to different sensors used in real-world well log application. These various features have different physical characteristics and can be used for different application purposes. With the consideration of generalization and scalability, we choose gamma ray (GR), density (DEN), acoustic (AC), spontaneous potential (SP), compensated neutron (CN), and resistivity (RT) as the original features for well log data representation. These features are summarized in Table 1.

The GR feature is obtained by measuring the intensity of gamma rays emitted during the decay process of naturally occurring radionuclides in the rock formation in the well. The GR feature can be used to judge lithology, stratigraphic comparison, and estimate shale content. The DEN feature is obtained by using an isotope gamma-ray source to radiate gamma rays to the formation, and then using a detector at a certain distance from the gamma source to measure the intensity of the gamma rays that reach the detector after being scattered and absorbed by the formation. The intensity of the scattered gamma rays received by the detector is related to the rock volume density of the formation. The SP feature is a measurement of the potential generated under the electrochemical action of the formation. The "positive", "negative" and amplitude of the

Table 1. The original features for well log data representation.

Original features	Physical characteristics and utilization
GR	Stratigraphic comparison
DEN	Identify lithology
AC	Calculate mineral content
SP	Determine the flooded layer
CN	Determine the gas layer
PT	Depth correction

spontaneous potential polarity are consistent with the relationship between the mud filtrate resistivity Rmf and the formation water resistivity Rw. When Rmf $\cong$ Rw, SP is almost flat; when Rmf $>$ Rw, SP is negative anomaly; when Rmf $<$ Rw, SP is positive anomaly in the permeable layer. The SP feature can be used to divide permeability stratum, determine the resistivity of formation water, determine the water-flooded layer, and study the sedimentary facies. The AC feature is based on the acoustic physical characteristics of the rock to measure the sound velocity of the formation. When the gas-bearing layer appears, the sound wave time difference appears to have cycle jumping phenomenon, or the logging value becomes larger. At large boreholes, the sonic jet difference will increase or jump. The CN feature is a thermal neutron logging using the dual source distance ratio method, which measures the thermal neutron flux caused by the neutron source along the well section. The CN feature can be used to determine the porosity of the formation, calculate the mineral content, and judge the gas layer by overlapping the compensation density curve. The RT feature is a logging curve that studies the electric field distribution in various media. In the measurement, an artificial electric field is first introduced into the medium. The distribution characteristics of this field are determined by the resistivity of the surrounding medium. Therefore, as long as the electric field distribution characteristics in various media are measured, the resistivity of the medium can be determined. The RT feature can be used to divide lithology profile, find the true resistivity of the rock formation, find the porosity of the formation, depth correction, and formation comparison.

3.2 Feature Representation Learning

The proposed autoencoder model is on the basis of deep neural network. It consists of multiple fully-connected (FC) layers. It is proposed to explore the intrinsic representation of well log data. The autoencoder model is trained by minimizing the reconstruction error between the original features and predicted features.

The number of encoder layers needs to be associated with the characteristics of deep neural networks. Increasing the number of FC layers can reduce network errors and improve accuracy. However, more FC layers will make the network much more complex and lead to the overfitting problem. It will also increase the training time of the deep neural network. Generally, increasing the number of neuron in FC layers can achieve a lower reconstruction error, while the training effect is easier to be achieved than that by increasing the number of hidden layers. For a neural network model without hidden layers, it is a linear or nonlinear regression model. According to the cross verification for setting the number of layers of the autoencoder, it can be obtained by comparing with multiple experiments based on well logs data.

The number of nodes in each hidden layer of the encoder needs to be combined with the characteristics of the neural network and the distribution of potential characteristics of the data. In autoencoder model, the choice of the number of neurons in FC is very important. It not only has a great impact on the performance of the established neural network model, but also is the potential cause of the overfitting during training process. However, there is no scientific and universal method of determination in theory. The calculation formulas for determining the number of neurons in FC layer in the literature are proposed for the case of any number of training samples. And most of these methods are used for the most unfavorable situation, which is difficult to be satisfied in general engineering practice and should not be adopted. In fact, the number of neurons in FC layer obtained by various calculation formulas sometimes differs several times or even hundreds of times. In order to ensure sufficiently high network performance and generalization ability, it is crucial to deal with the overfitting problem during training as much as possible. Therefore, the basic principle for determining the number of neurons in FC layers is to use as compact a structure as possible with the consideration for satisfying the accuracy requirements, i.e., using as less neurons as possible. Research shows that the number of neurons in FC layer is related to the size of the input and output in the specific layer. And it is also related to factors such as the complexity of the problem, the type of the transfer function, and the characteristics of the sample data. The number of training samples must be more than the connections of the neural network model. If the number of neurons is too large, the learning ability of the encoder will be improved. But if the number of neurons is too small, the encoder may not be able to learn any information. It is necessary to set a proper number of neurons for the encoder to ensure that the encoder learns effective information. According to experimental verification on well logs data, the number of neurons in each layer is chosen carefully and wisely.

Taking the above factors into consideration, the encoder network structure is set as four layers. The decoder structure is a symmetrical mirror of the encoder. The decoder and encoder structure are only the reverse relationship correspondingly. The network parameters are the settings of the corresponding layers of the encoder are the same. The structure of the autoencoder model is summarized in Table 2.

Due to complex environment of the well logging in real world application, signal noise and machine disturbing are inevitable during data collecting process. There are great amount of random noise, which will affect impact feature

Table 2. The structure of autoencoder model

Layer (type)	Structure
Input	Size: 6
FC Layer 1	Output number: 64, dropout: 0.5, activate: elu
FC Layer 2	Neuron number: 128, dropout: 0.5, activate: elu
FC Layer 3	Neuron number: 256, dropout: 0.5, activate: elu
FC Layer 4	Neuron number: 512, dropout: 0.5, activate: elu
FC Layer 5	Neuron number: 256, dropout: 0.5, activate: elu
FC Layer 6	Neuron number: 128, dropout: 0.5, activate: elu
FC Layer 7	Neuron number: 64, dropout: 0.5, activate: elu
Output	Size: 6

representation learning. Therefore, we use a more robust elu activation function for each neurons. In the different perspective view, the elu activation function can still propagate the gradient under any condition. The elu activation function can prevent the death of neurons and update connection weights, and speed up the convergence of the network.

The reconstruction error is measured between original features and predicted features. In our proposed method, we use the mean square error (MSE) as the loss function. The MSE loss function is defined as:

$$MSE = \frac{1}{m}\sum(y - y')^2, \tag{1}$$

where y is the input original features and y' is predicted features of the decoder, and m is the number of output. We use the Adam optimizer for gradient descent training, due to its stable performance and suitable for large-scale data scenarios.

3.3 Stratigraphic Interface Detection

In this section, we utilize the learned feature representation for well log interpretation application. Log interpretation is to analyze the classification information of each depth point according to the characteristics of the well logs data. Stratigraphic interface detection is the first and important application for in well log interpretation. Experts divide geology into more than 20 types of geological layers. Stratigraphic interface detection purpose is to associate the each well log data with specific geological type. It can then be formulated to use classifier to classify the well log data with different class label. Since linear SVM has the characteristics of fast classification speed and high classification accuracy, we choose linear SVM as the classifier.

Classification is a very important task in the field of data mining. The SVM classifier is a classification algorithm in machine learning. The purpose of SVM is to learn a classification function, such that the classifier can be used to predict unseen sample. The linear SVM classifies data by calculating a hyperplane, which minimizes the classification error and maximizes the classification interval. The classification interval is defined as:

$$L = max\frac{1}{2}||w||^2 - C\sum(y(w \cdot x + b) - 1), \tag{2}$$

where $w * x + b$ is the classification plane, and w is the normal vector of the classification plane. In Eq. 2, C is the penalty coefficient, C represents the degree of attention to the total error during the entire optimization process. A larger value of C leads to a higher the requirement for reducing the error.

4 Experiments

In order to evaluate the performance of proposed feature representation learning method, we choose a well logging data collected in real-world oilfield application. There are total 37761 samples with 8 geological type, containing 'J2z', 'J1y', 'J1f', 'chang1', 'chang2', 'chang3', 'chang4+5', and chang6'. We randomly select 70% of the data for the training and use the remaining data for the testing. We choose the principal component analysis (PCA) and locally linear embedding (LLE) as the baseline. We use PCA to preserve 3 dimension with 95% information on the basis of original features. We use kmeans method for clustering to evaluate the linear separability of feature representation.

4.1 Autoencoder Performance

We divide the input data into a testing set and a training set. We use the training set to train the autoencoder model, and then use the testing set to verify the performance of the network model. After 100 epochs of the training, the accuracy curve and loss curve are shown in Fig. 3.

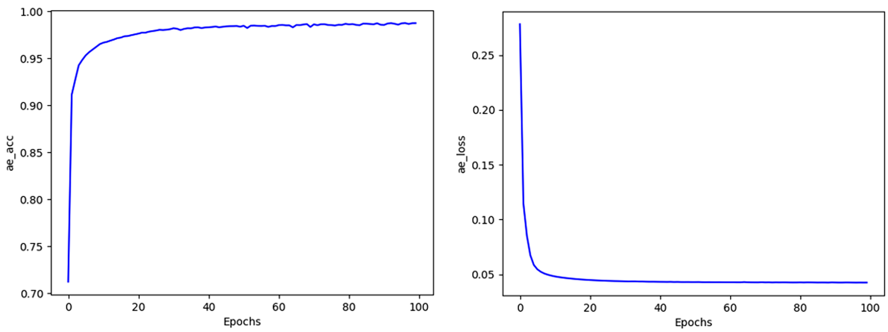

Fig. 3. The accuracy and loss of our proposed autoencoder model during the training process.

As shown in Fig. 3, the reconstruction error reduce rapidly and the training accuracy increases rapidly. After 20 epochs, the loss and accuracy rate remain stable. The accuracy rate achieve 99%, while the reconstruction error between the predicted features and the original features was smaller than 0.04. The learned 512-dimensional feature representation is supposed to contain all the information as that of the original features.

We selected 10000 samples and used the t-SNE [20] method to project the original features and our learned features to 2 dimensions plane for visualization. The result shown in Fig. 4.

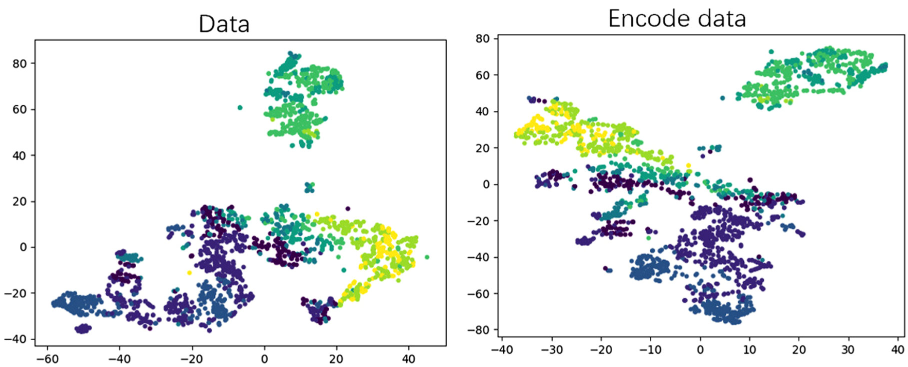

Fig. 4. Visualization of original features and our learned features. Different color indicate different class of geological type.

As the left part shown in Fig. 4, we can see that the samples formulated by original features from different classes are coupled together. It is not easy for a classifier to make classification. As the right part shown in Fig. 4, the learned features reduce the coupling and increases the intra-class variance. It will be easy to achieve better classification using a linear classifier. Therefore, the learned feature representation shows better separability and feature expression.

4.2 Classification Evaluation

In order to evaluate the feature representation, we use linear SVM for stratigraphic interface detection to evaluate the performance of the feature representation. Since the penalty coefficient C of linear SVM has an impact on the classification effect, we use different penalty coefficients during the experiments. We compare classification performance between our feature representation with that of the original features, PCA, and LLE. The classification results are summarized in Table 3.

As shown in Table 3, when the penalty coefficient increases, the classification accuracy of SVM also increases. When the penalty coefficient is set as 100, the classification accuracy of the original features achieves 74.7%, and our proposed

Table 3. The classification results using different C

C	Original	PCA	LLE	Ours
0.0001	0.530	0.485	0.243	0.675
0.001	0.608	0.521	0.243	0.768
0.01	0.686	0.557	0.243	0.865
1	0.739	0.558	0.475	0.979
10	0.744	0.558	0.459	0.994
100	0.747	0.558	0.457	0.998

feature representation is able to achieve 99.8%. Compared with the original features, the classification accuracy has been greatly improved. The experiments indicate that our proposed feature representation has better separability.

When the penalty coefficient increases to 1, the improvement of SVM classification accuracy has stabilized. We summary the statistics of the classification using our proposed feature representation in Table 4:

Table 4. The statistics of classification using our proposed feature representation with different C.

Label	Number	C = 0.0001	C = 1	C = 100
0	1012	0.544	0.934	0.996
1	2755	0.788	0.997	0.998
2	1682	0.758	0.998	1
3	815	0.744	0.963	0.996
4	1372	0.506	0.964	1
5	1700	0.635	0.986	0.996
6	1422	0.559	0.969	0.997
7	560	0.425	0.962	0.997
Total	11318	0.675	0.979	0.998

4.3 Clustering Evaluation

Clustering is one of the commonly used algorithms in machine learning, and the clustering evaluation index can intuitively reflect the feature representation ability. We use the Kmeans algorithm for clustering the data processed by PCA, LLE, and our proposed method. We use adjusting rand index (ARI), adjusting mutual information (AMI), V-measure, homogeneity, completeness, silhouette, and davies bouldin index for comparison [18,19]. The results are summarized in Table 5.

Table 5. Clustering evaluation index comparison

Evaluation indexl	Data	PCA	LLE	Ours
ARI	0.292	0.282	0.239	0.306
AMI	0.395	0.399	0.327	0.430
V-measure	0.396	0.400	0.328	0.430
Homogeneity	0.387	0.396	0.298	0.431
Completeness	0.405	0.404	0.366	0.429
Davies bouldin index	1.081	0.968	0.822	1.144
Silhouette	0.297	0.316	0.305	0.320

As shown in Table 5, we can see that the our proposed feature representation is better than that of the original features, PCA, and LLE through all clustering evaluation. Compared with traditional dimensionality reduction methods, our proposed method has better feature expression capabilities.

5 Conclusion

This paper proposed a data processing method based on autoencoder. According to the characteristics of the well logs data and the problems in the logging process, a feature coding network is set up to upgrade well logs data. We selected a well containing 37761 data, and used linear SVM classification and clustering algorithm to verify the effect of the proposed method. And compared with the commonly used feature extraction methods PCA and LLE to verify the effectiveness of the encoding. The experimental results demonstrate that the classification accuracy by using learned feature representation increase to 99.8% compared with that of 74.6% by using original feature representation,so the encoded well logs data has better separability. From the evaluation index of KMeans algorithm, the encoded data has better feature expression ability than the original data.

Stratigraphic division is only the first step in logging interpretation. Logging interpretation includes tasks such as lithology identification, reservoir division. In the future, we will continue to explore the application of this feature representation method we proposed in these tasks. Provide more methods to improve the efficiency and accuracy of logging interpretation.

References

1. Wang, Y., Liu, C., Yuan, X.: Stacked locality preserving autoencoder for feature extraction and its application for industrial process data modeling. Chemometr. Intell. Lab. Syst. **203**, 104086 (2020)
2. Mathematics - Geometry; Reports from China University of Petroleum (East China) Advance Knowledge in Geometry (Numerical Study of Pore Structure Effects On Acoustic Logging data In the Borehole Environment). J. Math. (2020)

3. Jo, S., Sohng, W., Lee, H., Chung, H.: Evaluation of an autoencoder as a feature extraction tool for near-infrared spectroscopic discriminant analysis. Food Chemistry **331**, (2020)
4. Zhang, L., et al.: Research on image transmission mechanism through a multimode fiber based on principal component analysis. Opt. Lasers Eng. **134**, (2020)
5. Nagarajan, S., Nettimi, S.S.S., Kumar, L.S.. Nath, M.K., Kanhe, A.: Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales. Digit. Signal Process. **104** (2020)
6. Li Kai, H., Shaohua, Z.F., et al.: Application of relief F algorithm in attribute optimization. Petrol. Geophys. Explor. **53**(S2), 230–234 (2018)
7. Applied radiation research; new findings in applied radiation research described from China University of petroleum (neutron transport correction and density calculation in the neutron-gamma density logging). Sci. Lett. (2019)
8. Wu, S., Liu, Y., Wang, J., Li, Q.: Sentiment analysis method based on Kmeans and online transfer learning. Comput. Mater. Continua **60**(3), 1207–1222 (2019)
9. Mei, H., Hanming, G., Huan, W.: Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir. J. Pet. Sci. Eng. (2020, prepublish)
10. Szabó, N.P., Dobróka, M.: Interval inversion as innovative well log interpretation tool for evaluating organic-rich shale formations. J. Petrol. Sci. Eng. **186**(C) (2020)
11. Geophysics; Research Conducted at National Research Institute Has Provided New Information about Geophysics (Well logging interpretation methodology for carbonate formation fracture system properties determination). Sci. Lett. (2020)
12. Davydycheva, S., Kaminsky, A.: Triaxial induction logging: new interpretation method for biaxial anisotropic formations–Part 1. Soc. Explor. Geophys. Am. Assoc. Petrol. Geol. **4**(2), 151–164 (2016)
13. Viggen, E.M., Merciu, I.A., Løvstakken, L., Måsøy, S.-E.: Automatic interpretation of cement evaluation logs from cased boreholes using supervised deep neural networks. J. Petrol. Sci. Eng. (2020)
14. Tan, M., Bai, Y., Zhang, H., Li, G., Wei, X., Wang, A.: Fluid typing in tight sandstone from wireline logs using classification committee machine. Fuel **271** (2020)
15. Rodríguez-Hoyos, A., Rebollo-Monedero, D., Estrada-Jiménez, J., Forné, J., Urquiza-Aguiar, L.: Preserving empirical data utility in k -anonymous microaggregation via linear discriminant analysis. Eng. Appl. Artif. Intell. **94** (2020)
16. Wang, Y., Liu, M., Bao, Z., Zhang, S.: Stacked sparse autoencoder with PCA and SVM for data-based line trip fault diagnosis in power systems. Neural Comput. Appl. **31**(10), 6719–6731 (2018). https://doi.org/10.1007/s00521-018-3490-5
17. Tao, Z.: Application of BP Neural Network in Logging Interpretation. Northwest University (2010)
18. Osamor, I.P., Osamor, V.C.: OsamorSoft: clustering index for comparison and quality validation in high throughput dataset. J. Big Data **7**(1), 1–13 (2020). https://doi.org/10.1186/s40537-020-00325-6
19. Bhatia, N., Sojan, J.M., Simonovic, S., Srivastav, R.: Role of cluster validity indices in delineation of precipitation regions. Water **12**(5) (2020)
20. Nicola, P., Lelieveldt Boudewijn, P.F., Van Der Maaten, L., Thomas, H., Elmar, E., Anna, V.: Approximated and user steerable tSNE for progressive visual analytics. IEEE Trans. Vis. Comput. Graph. **23**(7) (2017)

21. Xu, S.: The fuzzy comprehensive evaluation (FCE) and the principal component analysis (PCA) model simulation and its applications in water quality assessment of Nansi Lake Basin, China. Environ. Eng. Res. 26(2) (2021)
22. Liang, L., et al.: Reservoir characteristics and logging evaluation of gas-bearing mudstone in the south of North China plain. Sci. Rep. **10**(1) (2020)