Artificial Intelligence in Geosciences 1 (2020) 24–30

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences

Seismic labeled data expansion using variational autoencoders

Kunhong Li [a,b,*], Song Chen [a], Guangmin Hu, Ph.D [a,b]

[a] *School of Resources and Environment and Center for Information Geoscience, University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China*
[b] *Center for Information Geoscience, UESTC, Chengdu, 611731, China*

ARTICLE INFO

Keywords:
Deep learning
Variational autoencoders
Data expansion

ABSTRACT

Supervised machine learning algorithms have been widely used in seismic exploration processing, but the lack of labeled examples complicates its application. Therefore, we propose a seismic labeled data expansion method based on deep variational Autoencoders (VAE), which are made of neural networks and contains two parts-Encoder and Decoder. Lack of training samples leads to overfitting of the network. We training the VAE with whole seismic data, which is a data-driven process and greatly alleviates the risk of overfitting. The Encoder captures the ability to map the seismic waveform $\mathbf{Y}$ to latent deep features $\mathbf{z}$, and the Decoder captures the ability to reconstruct high-dimensional waveform $\widehat{\mathbf{Y}}$ from latent deep features $\mathbf{z}$. Later, we put the labeled seismic data into Encoders and get the latent deep features. We can easily use gaussian mixture model to fit the deep feature distribution of each class labeled data. We resample a mass of expansion deep features $\mathbf{z}^*$ according to the Gaussian mixture model, and put the expansion deep features into the decoder to generate expansion seismic data. The experiments in synthetic and real data show that our method alleviates the problem of lacking labeled seismic data for supervised seismic facies analysis.

1. Introduction

Supervised machine learning algorithms have been widely used for seismic interpretation in recent years, such as fault detection (Wu et al., 2019), seismic facies (Wrona et al., 2018). A key to successful application of supervised machine learning is labeled training. Seismic label mainly come from well loggings data and interpreters, both of which are limited.

Manual interpretations are one of the most direct and intuitively simplest ways to increase labeled samples. Therefore, many seismic interpretation methods use the results of manual interpretation as training samples. However, manual interpretations require experience and take up only a small percentage of the whole seismic data set. In addition, labeled sample data can be extended by forward simulations based on human prior knowledge. Above methods still require expert knowledge as well as significant time, so we propose a data-driven method for the seismic labeled data expansion.

The proposed method is based on variational autoencoders (VAE), which are unsupervised models widely used for attribute extraction. A VAE is made of multilayer neural networks, which consist of one encoder and one decoder. The encoder codes the high-dimensional seismic data into the low-dimensional deep feature space, and then the decoder maps the deep features to reconstruct the seismic waveform. The goal of VAE is to minimize the difference between the input waveform and the reconstructed waveform. Therefore, the VAE does not need labeled data and is

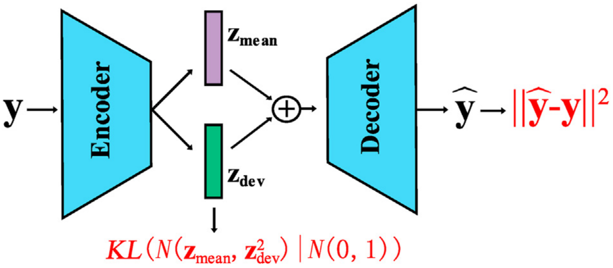

y
Encoder
$\mathbf{z}_{mean}$
$\mathbf{z}_{dev}$
$\oplus$
Decoder
$\widehat{\mathbf{y}}$
$||\widehat{\mathbf{y}}-\mathbf{y}||^2$
$KL(N(\mathbf{z}_{mean}, \mathbf{z}_{dev}^2)\,|\,N(0,1))$

Fig. 1. The structure of VAE. The red font is the loss function. The $KL(\cdot|\cdot)$ is the KL divergence. The $||\cdot||$ is the 2-norm and the $N(\cdot)$ is the gaussian distribution. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

[*] Corresponding author. School of Resources and Environment and Center for Information Geoscience, University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China.

E-mail addresses: lkhmoran@gmail.com (K. Li), 2547546252@qq.com (S. Chen), hgm@uestc.edu.cn (G. Hu).

https://doi.org/10.1016/j.aiig.2020.12.002
Received 31 August 2020; Received in revised form 20 November 2020; Accepted 12 December 2020
Available online 13 January 2021

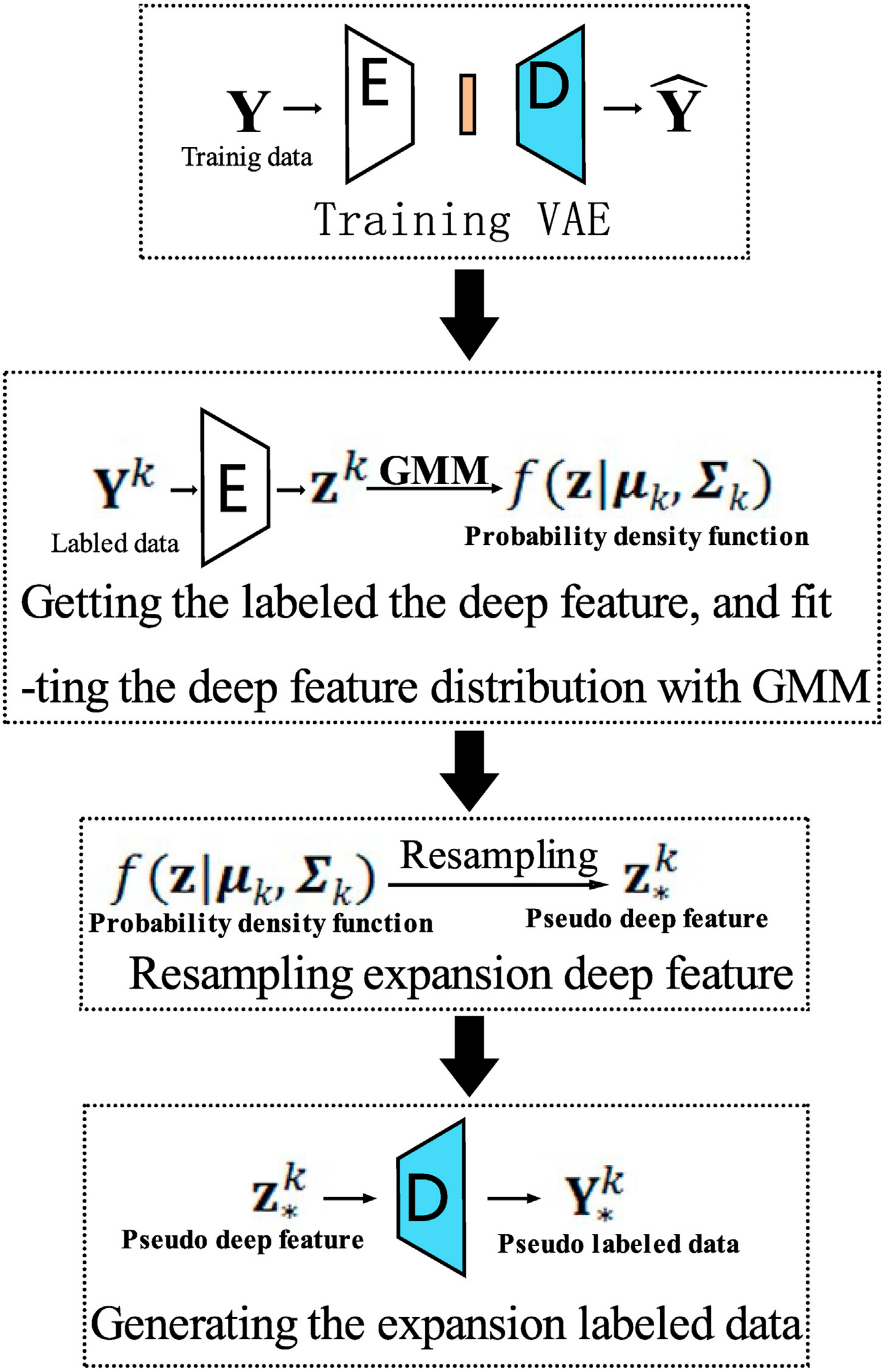

Y
Trainig data
E
D
Ŷ
Training VAE
Y^k
Labled data
E
z^k
GMM
$f(z|\mu_k, \Sigma_k)$
Probability density function
Getting the labeled the deep feature, and fit -ting the deep feature distribution with GMM
$f(z|\mu_k, \Sigma_k)$
Probability density function
Resampling
z_*^k
Pseudo deep feature
Resampling expansion deep feature
z_*^k
Pseudo deep feature
D
Y_*^k
Pseudo labeled data
Generating the expansion labeled data

Fig. 2. The workflow of the proposed method.

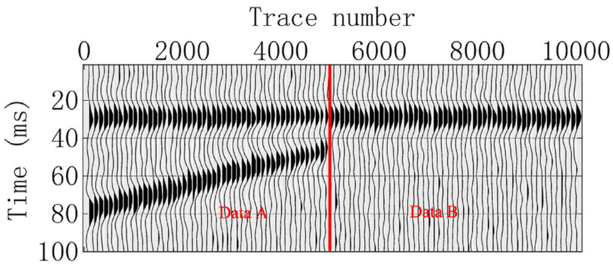
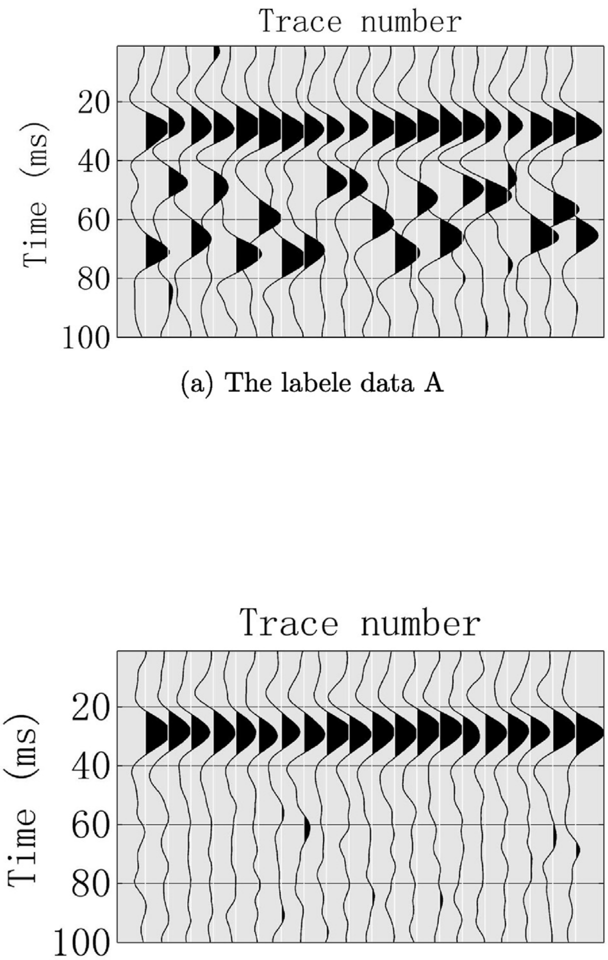


Fig. 3. The synthetic data.



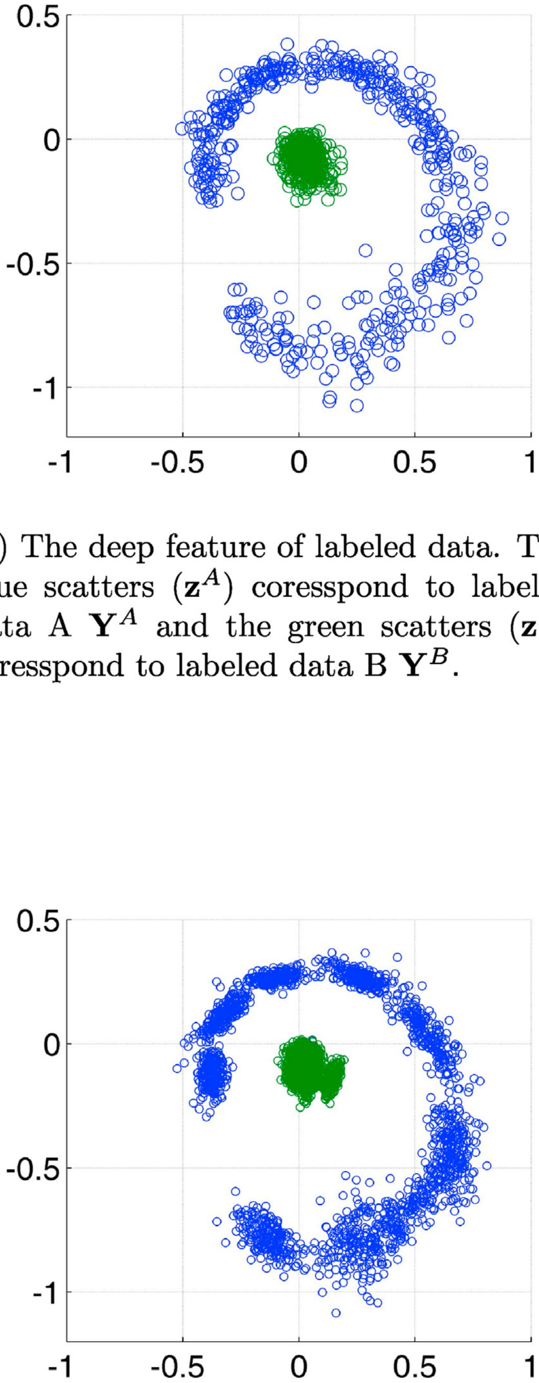
(a) The labele data A

(b) The labele data B

Fig. 4. The labeled data. (a) The labele data A. (b)The labele data B.

an unsupervised manner with VAE. Massive seismic data can be used as training data, which can greatly alleviates the risk of over-fitting.

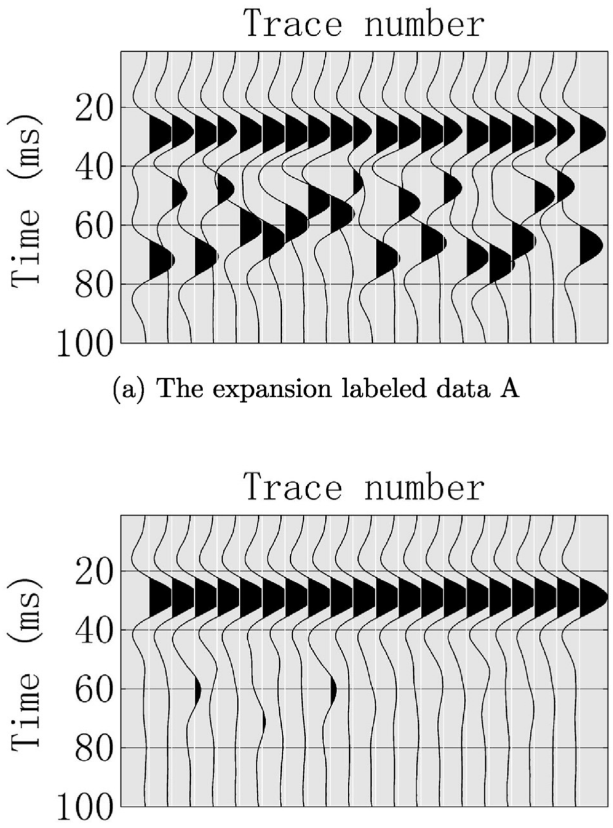
The VAE captures the deep feature distribution of seismic data when it converges. We put the labeled samples into the decoder, and obtain the deep feature distribution of each class. It is easy to fit the deep feature distribution with the gaussian mixture model (GMM) for each class. It should be noted that each class of data following one gaussian mixture



(a) The deep feature of labeled data. The blue scatters ($\mathbf{z}^A$) coresspond to labeled data A $\mathbf{Y}^A$ and the green scatters ($\mathbf{z}^B$) coresspond to labeled data B $\mathbf{Y}^B$.

(b) The expansion deep feature. The blue scatters ($\mathbf{z}_*^A$) coresspond to expansion labeled data A $\mathbf{Y}_*^A$ and the green scatters ($\mathbf{z}_*^B$) coresspond to expansion labeled data B $\mathbf{Y}_*^B$.

Fig. 5. The deep feature. (a) The deep feature of labeled data. (b)The expansion deep feature.



(a) The expansion labeled data A

(b) The expansion labeled data B

Fig. 6. The expansion labeled data. (a) The expansion labeled data A. (b)The expansion labeled data B.

distribution rather all the class data follow a gaussian mixture distribution. In short, each class deep feature has an explicit gaussian distribution function. Subsequently, we get expansion deep feature by using these functions. Finally, we put the expansion deep features through the decoder and get expansion labeled data.

We proposed a semi-supervised labeled data expansion method. One of the main reasons for network overfitting is the lack of training samples. This method uses whole seismic data for the training of feature extraction and alleviates the problem of over-fitting. We use this method for synthetic and real data and the results demonstrate that the generated expansion labeled data enhance the seismic waveform classification performance. In the synthetic data experiment, the classification accuracy is improved by 20 % with the proposed method when SNR is 3.

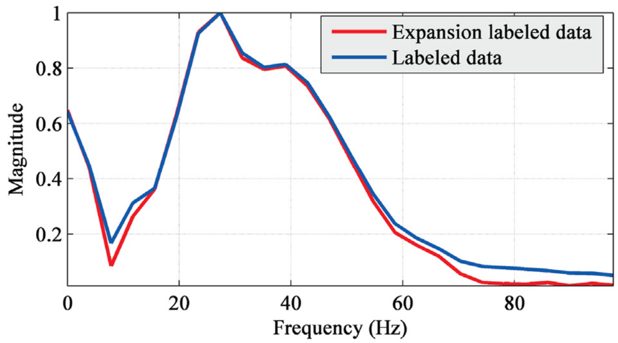
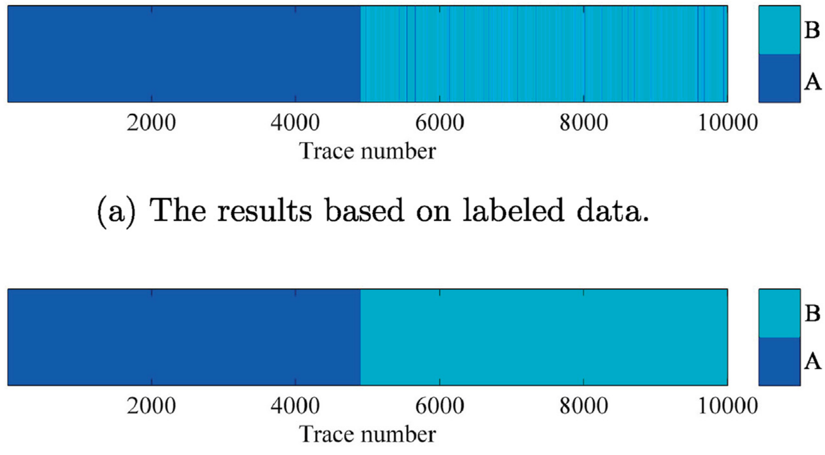


Fig. 7. Spectrum of synthetic data. The labeled data and expansion labeled data almost have the same spectrum curve.

2. Theory

2.1. *Variational autoencoder*

The VAE is defined by a neural network, which consists of two parts, the encoder E and the decoder D. The encoder codes the seismic signal $\mathbf{y} \in \mathbb{R}^M$ into deep feature mean $\mathbf{z}_{mean}$ and deviation $\mathbf{z}_{dev}$, and the deep feature $\mathbf{z}$ equals $\mathbf{z}_{mean}$ plus $\mathbf{z}_{dev}$, i.e. $\mathbf{z} = \mathbf{z}_{mean} + \mathbf{z}_{dev}$. Then the decoder D is used to decode the feature $\mathbf{z}$ to generate the seismic signal $\widehat{\mathbf{y}}$, i.e. $\widehat{\mathbf{y}} = D(\mathbf{z})$. The purpose of VAE is to fit the generated $\widehat{\mathbf{y}}$ to the input $\mathbf{y}$, while features $\mathbf{z}$ need to follow the prior gaussian distribution. Therefore, the loss function of the VAE includes two parts, one is the reconstruction error, which is measured by the Euclidean distance. Another part of the loss function is about the feature $\mathbf{z}$, and the KL divergence be used to measure the differences between $\mathbf{z}$ and prior gaussian distributions. Hence the VAE loss function is defined as follows:



(a) The results based on labeled data.

(b) The results based on labeled data and expansion labeled data.

Fig. 8. The SVM waveform classification results with same parameters (kernel function is radial basis function and sigma is 3). (a) The results based on labeled data. (b)The results based on labeled data and expansion labeled data.

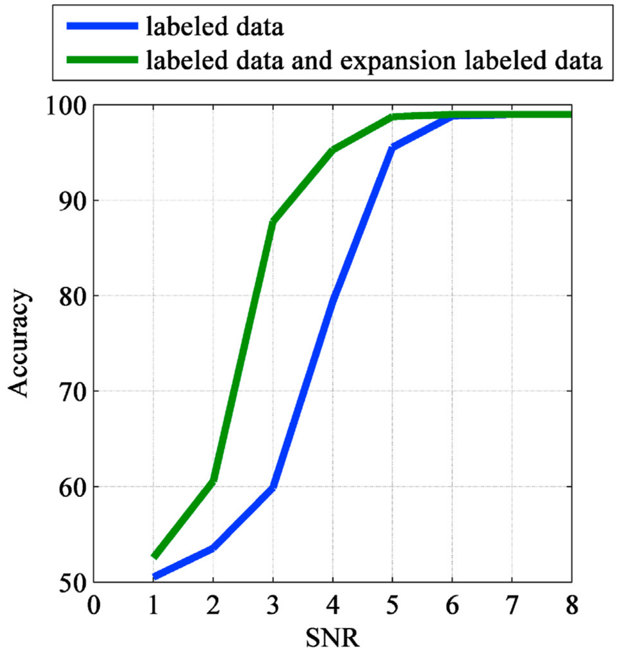


Fig. 9. Classification accuracy vs SNR.

$$L_{VAE} = \mathbf{y}log(\mathbf{y}) + (1-\mathbf{y})log(1-\mathbf{y}) + \frac{1}{2}\left(-log\left(\mathbf{z}_{dev}^2\right) + \mathbf{z}_{dev}^2 + \mathbf{z}_{dev}^2 - 1\right) \tag{1}$$

Due to space constraints, please refer to (Doersch, 2016) for the details of Eq. (1). The structure of VAE is shown in Fig. (1). We use gradient descent to solve Eq. (1).

2.2. Gaussian mixed model

The probability density function of N-dimensional Gaussian distribution is defined as follow:

$$N\left(\mathbf{y}\middle|\mu,\Sigma\right) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\Sigma^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)\right) \tag{2}$$

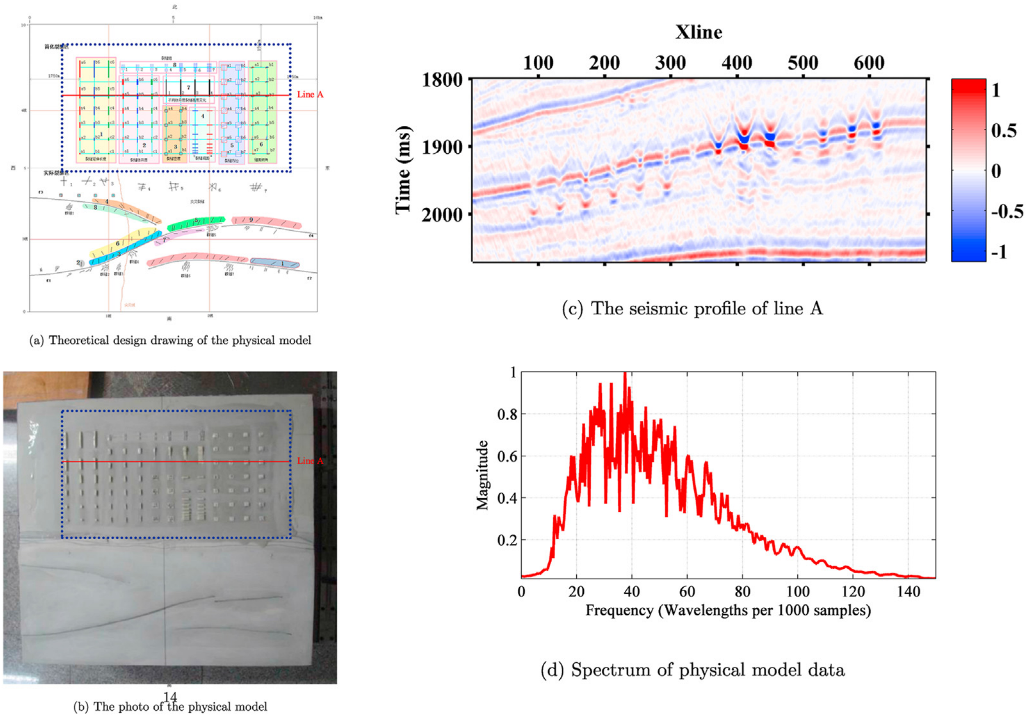
where $\mu \in \mathbb{R}$ is the mean vector and $\Sigma \in \mathbb{R}^{N\times N}$ is the covariance matrix. A distribution consisting of K Gaussian distributions is called a Gaussian mixed Model (GMM), and its probability density function is:

$$f\left(y|\Theta\right) = \sum_{k=1}^{K} \alpha_k N(y|\mu_k, \Sigma_k) \tag{3}$$

where α_k is the weight coefficient and $\sum_{k=1}^{K}\alpha_k = 1$. The Θ represents all the parameters $(\alpha_k, \mu_k, \Sigma_k)$. The expectation-maximization algorithm (Dempster et al., 1977) can be used to solve the Eq. (3). We chose GMM to fit distribution of the deep feature **z** because of that GMM can theoretically fits the probability distribution of any shape.

2.3. Labeled samples expansion

The workflow of the proposed method is shown in Fig. 2. There are whole seismic data set $\mathbf{Y} \in \mathbb{R}^{M\times m}$ and the labeled set $\mathbf{Y}^k \in \mathbb{R}^{M\times m_k}$, where M is the dimension, m is the number of samples, k is the index of category and m_k is the number of k category labeled samples. The proposed method has a total four steps (2). The first step is training the VAE with the whole data set **Y**. The Encoder of VAE will captures distribution of the deep feature **z** and the Decoder of VAE has the ability that generating waveform from the deep feature **z**. The second step is putting labeled samples $\mathbf{Y}^k$ into the Encoder and getting the corresponding the deep feature $\mathbf{z}^k$, which fits the each class deep feature distribution with GMM. We get the probability density function $f(z|\mu_k, \Sigma_k)$ for each class. The third step is resampling pseudo deep feature z_k^k via $f(z|\mu_k,\Sigma_k)$, and the final step is to put the z_*^k into



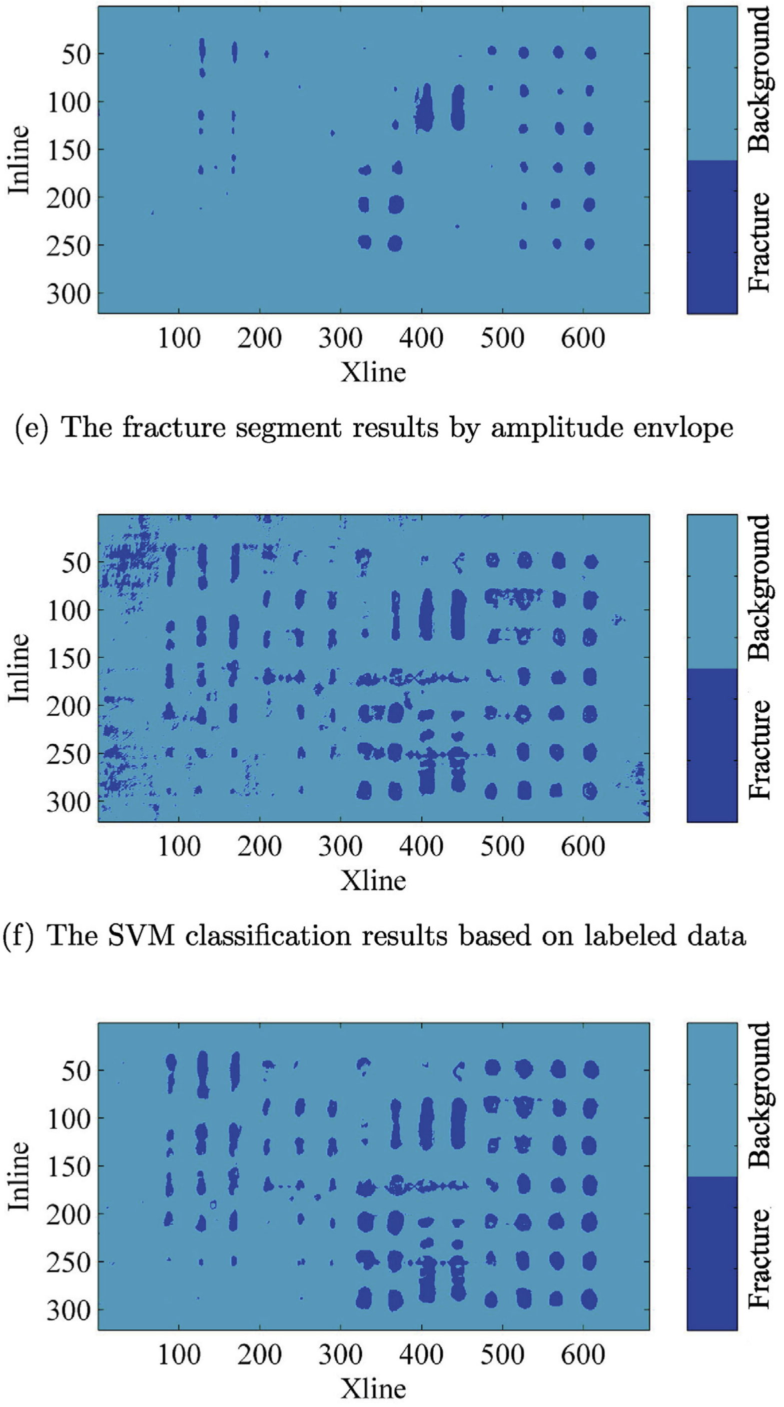
(a) Theoretical design drawing of the physical model

(b) The photo of the physical model

(c) The seismic profile of line A

(d) Spectrum of physical model data

Fig. 10. The information about the physical model. (a) Theoretical design drawing of the physical model. (b)The photo of the physical model. (c) The seismic profile of line A. d) Spectrum of physical model data.



(e) The fracture segment results by amplitude envlope

(f) The SVM classification results based on labeled data

(g) The SVM classification results based on labeled data and expansion data.

Fig. 11. The fracture identification results. (e) The fracture segment results by amplitude envelope. (f)The SVM classification results based on labeled data. (g) The SVM classification results based on labeled data and expansion labeled data.

the Decoder to generate the pseudo labeled sample $\mathbf{Y}_*^k$.

3. Experiments

3.1. Synthetic data

We design a synthetic data set shown in the Fig. 3. The synthetic data has total 10,000 traces with added band-limited gaussian noise (SNR = 3). The red line separates the synthetic data into two categories, the left is data A and the right is data B. We take all the data as the training set and randomly select 500 traces from data A and B as labeled data respectively, i.e. $\mathbf{Y}^A$ and $\mathbf{Y}^B$. The Fig. 4 shows the labeled data. We use the proposed method to extend the labeled data. For better visualization, we set the deep latent feature $\mathbf{z}$ to 2 dimensions. The Fig. 5 shows the deep latent features. The blue scatters ($\mathbf{z}^A$) correspond to the labeled data A ($\mathbf{Y}^A$) and the green scatters ($\mathbf{z}^B$) correspond to labeled data B ($\mathbf{Y}^B$). We fit the deep latent feature with GMM and resample 2000 expansion deep features for each class data ($\mathbf{z}_*^A$, $\mathbf{z}_*^B$), which are shown in Fig. 5b. We can see that the expansion deep feature ($\mathbf{z}_*^A$,$\mathbf{z}_*^B$) is well in line with the labeled deep feature distribution and greatly expands the volume. Finally, we get the expansion labeled data $\mathbf{Y}_*^A$, $\mathbf{Y}_*^B$ that is shown in Fig. 6. Fig. 7 shows the spectrum curves of labeled data and expansion labeled data. The generated data has same character with the labeled data, and meanwhile has less noise. The Fig. 8 shows the seismic waveform classification results via support vector machine (SVM). The Fig. 8a is based on labeled data ($\mathbf{Y}^A$ and $\mathbf{Y}^B$) and the Fig. 8b is based on the labeled data ($\mathbf{Y}^A$ and $\mathbf{Y}^B$) and expansion data ($\mathbf{Y}_*^A$ and $\mathbf{Y}_*^B$). For fairness, we set the same SVM parameters for both datasets, (kernel function is radial basis function and sigma is 3). Fig. 9 shows the accuracy of classification under different SNR. Obviously, the result based on the labeled data ($\mathbf{Y}^A$ and $\mathbf{Y}^B$) and expansion data ($\mathbf{Y}_*^A$ and $\mathbf{Y}_*^B$) has better accuracy and robustness, which indicates the expansion data improves the performance of classification.

3.2. Real data

We applied the proposed methods on a physical model data, which are provided by China national petroleum corporation key laboratory of geophysical exploration. The Fig. 10 shows the information about the physical model. Fig. 10a is the theoretical design drawing of the physical model, and Fig. 10b is the photo of the physical model. The main purpose of designing this physical model is to study fractures and fault. The area enclosed by the blue dotted line is the fractures zone. Fractures with different sizes, orientations, dip angles and densities are distributed in this zone. The major frequency of seismic data is at 40 Hz and the interval of seismic data is 1 ms.

Simply, we just use the data of the fractures area. There is a red line A cross the fractures area, and the seismic profile of line A is shown in Fig. 10c. We cut the waveform data along a horizon with 40 ms window. In order to get labeled data, we calculate the amplitude envelope attribute and then segment fracture via a threshold, as shown in Fig. 11e. We take 2000 samples from fracture and background respectively as labeled data, and then expand the label data using the proposed method. Based on these labeled and expansion data, we use SVM to classify seismic waveforms. For fairness, we set the same SVM parameters for both datasets, (kernel function is radial basis function and sigma is 5). The Fig. 11 shows the results of classification. The Fig. 11f is based on the labeled data and Fig. 11g is based on the labeled data and expansion data. The classification results based on expansion samples and sample data are closer to the real fracture distribution, which indicates that expansion samples play a role in improving the classification results to some extent.

4. Conclusion

We proposed a novel data-driven, semi-supervised label expansion method. We use whole survey seismic data as the training data for VAE, which provides enough training data and alleviates the risk of overfitting. The encoder of VAE projects seismic data into deep features, and the decoder of VAE generate seismic data from the deep features. We use GMM to extract deep feature distributions of labeled data, and then resample a large number of deep features. In other words, the proposed method extend labeled data in the deep feature space. We test our method via seismic waveform classification. The results of theoretical and practical data demonstrate that expansion data can well enhance the classification results.

Declaration of competing interest

The manuscript has not been published before and is not being considered for publication elsewhere. All authors have contributed to the creation of this manuscript for important intellectual content and read and approved the final manuscript. We declare there is no conflict of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (41804126, 41604107). We would like to thank China national petroleum corporation key laboratory of geophysical exploration for providing the physical model data.

References

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Stat. Soc. B 39, 1–22.

Doersch, C., 2016. Tutorial on Variational Autoencoders arXiv preprint arXiv: 1606.05908.

Wrona, T., Pan, I., Gawthorpe, R.L., Fossen, H., 2018. Seismic facies analysis using machine learning. Geophysics 83, O83–O95.

Wu, X., Liang, L., Shi, Y., Fomel, S., 2019. Faultseg3d: using synthetic data sets to train an end-to-end convolutional neural network for 3d seismic fault segmentation. Geophysics 84, IM35–IM45.