International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 5 Issue 8 ǁ August. 2017 ǁ PP. 42-48

Power Transformer Fault Diagnosis based on Deep Learning

Liu Bingyao[1], Wu Leibin[2]
1,2(College of Mechanical Engineering, Shanghai University of Engineering Science, China)
Correspondence author: Liu Bingyao

ABSTRACT: As a new method in machine learning, deep learning has achieved brilliant results in every field with its powerful self-feature extraction capacity in the last few years. Power transformer is the core of energy conversion and transmission in power grid, so that to learn its fault diagnosis methods is of important significance for detecting the faults earlier and making the system safer. Because the fault mechanism of the power transformer is complex, the thesis studies its fault diagnosis methods mainly via the vibration signal of power transformer, partial discharge signal and the gas content dissolved in the oil dissolved on the basis of the deep learning theory. The vibration signal is non-stationary and time-varying. If analyzed by traditional signal analysis technology, the primary function is fixed, so that its time-varying characteristics cannot be reflected. Therefore, the thesis tries to analyze the vibration signals of the power transformer via the local wave method. By recognizing the vibration modes within the signals, we can better discover the features of the vibration signal in case of faults, which can facilitate the fault diagnosis later. The sparse auto encoder can abstract data characteristics effectively, to make classificationeasier, which has a stronger optimization capacity compared to the traditional BP neutral network. When there is a failure in the power transformer, the online monitoring data of the oil chromatography shows no label at all, so that traditional fault diagnosis methods usually cannot take full use of such label-free fault samples. The thesis adopts the new CDLNN method to diagnose the faults in the power transformer, so that the shortcoming that BP cannot utilizes label-free samples can be overcome and, the research finds out that it has a stronger learning capacity.

Keywords: Deep Learning Neutral Network, Eigenvalue Extraction, Power Transformation, Sparse Auto Encoder

I. INTRODUCTION

As a deep machine learning method, the deep learning neutral network (DLNN) is qualified enough to extract features from samples and transform such features and, it has a strong learning ability, thus becoming a hot topic in domestic and foreign researches in recent years. TheDLNN adopts the non-supervision machine learning method in training, so that it can uses a great many of label-free samples to finish the pre-training process of the model, optimize model parameter, and improve the accuracy rate of the model in classification. Nowadays, it has been successfully applied in speech recognition, target recognition and natural language processing, yet its application in fault diagnosis of the power transformer has just started. Because the fault mechanism of the power transformer is complex, the thesis studies its fault diagnosis methods mainly via the vibration signal of power transformer, partial discharge signal and the gas content dissolved in the oil dissolved on the basis of the deep learning theory. The vibration signal is non-stationary and time-varying. If analyzed by traditional signal analysis technology, the primary function is fixed, so that its time-varying characteristics cannot be reflected. Therefore, the thesis tries to analyze the vibration signals of the power transformer via the local wave method. By recognizing the vibration modes within the signals, we can better discover the features of the vibration signal in case of faults, which can facilitate the fault diagnosis later. The sparse auto encoder can abstract data characteristics effectively, to make classificationeasier, which has a stronger optimization capacity compared to the traditional BP neutral network. When there is a failure in the power transformer, the online monitoring data of the oil chromatography shows no label at all, so that traditional fault diagnosis methods usually cannot take full use of such label-free fault samples. The thesis adopts the new CDLNN method to diagnose the faults in the power transformer, so that the shortcoming that BP cannot utilizes label-free samples can be overcome and, the research finds out that it has a stronger learning capacity. CDLNN adopts a semi-supervision machine learning method, which has a strong learning ability, can diagnose the probability of all running statuses of a power transformer and, provide more reference information for working staffs to decide whether to overhaul the power transformer or not.

1. DLNN algorithm

Deep learning neutral network (DLNN) is a deep machine learning method put forward by Professor Hinton in 2006. It is qualified enough to extract features from samples and transform such features and, it has a strong learning ability, thus becoming a hot topic in domestic and foreign researches in recent years. DLNN can be understood simply as a neutral network with multiple hidden layers, which discovers internal properties of

data by feature transformation or feature extraction, to make the classification easier and more accurate. DLNN method mainly includes auto encoder (AE), restricted Boltzmann machine (RBM) and convolutional neutral network (CNN). Among them, CNN is mainly used for image processing, rather than fault diagnosis of the power transformation, so that the author will not make more introduction to it later.

1.1 Sparse auto encoder

Sparse encoding was firstly put forward by OLSHAUSEN et.al., and it was used in the non-supervision calculation to simulate the perceptual learning of mankind. As an extension of auto encoder, sparse auto encoder uses the sparse encoding idea, introduces the sparse penalty item on auto encoder so that relatively concise and sparse data feature can be learnt under sparse constraint conditions, in order to better express the input data.The automatic encoder is a symmetrical three-layer neural network, which encodes the input data through the hidden layer, and then reconstructs the input data from the hidden layer to minimize the reconstruction error, and get best data hidden layer expression. According to the course notes of Andrew Ng, a professor from Stanford University, the theory of automatic encoders is briefly described below. A basic AE can be considered as a 3-layer neural network in which the output layer is of the same size as the input layer, as shown in Figure 1. Because there are some inherent problems in the auto coder. For example, make a simple copy and memory of input layer and deliver it to the hidden layer, which can perfectly recover the original input data, but such an automatic encoding cannot make any meaningful feature extraction. Therefore, the sparse automatic encoder improves the performance of the traditional automatic encoder, and therefore has more practical significance.

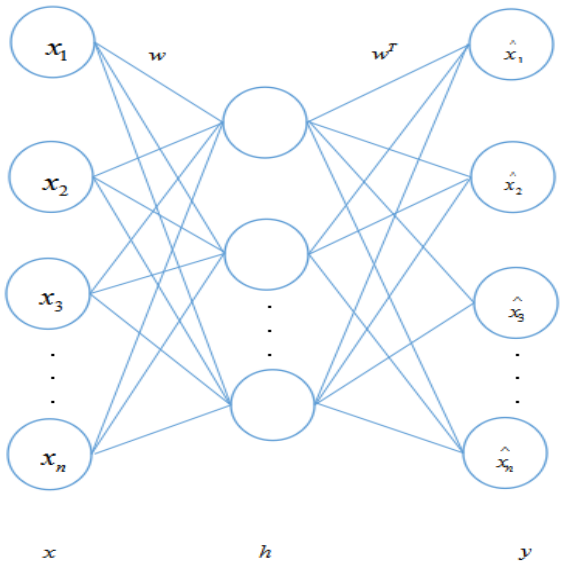


Fig. 1 AE structure

The collected data reflecting the vibration of the power transformer $y(i)=\{x_1,x_2,x_3,...,x_k\}$, will be reconstructed as $N\times M$ data array $\{x(1),x(2),x(3),...,x(N)\},x(i)\in R^M$. Namely, select N groups M-long data from each classification as the experiment data; set the input matrix constituted by the vibration data group as ***X***, the encoding process of the sparse automatic encoder toward the input data is described as below. Firstly, a three-layer neural network is set up: input layer, hidden layer and output layer, and sigmoid function is chosen as neuron activation function. For unlabeled input matrix ***X***, hope it can learn the feature expression of the hidden layer $h(X,W,b)=\sigma(WX+b)$ so that it can approximate or reconstruct the input data ***X***. The sparse penalty term is added to the cost function of the encoder, and it is eventually expected to control the number of activations of the hidden neurons. Normally, if the output of a neuron is approximately 1, it can be thought that the neuron is "active" and, conversely, the neuron is considered to be "inactive". One of the tasks of a sparse encoder is to make these neurons inactive at most times. Suppose $a_j(x)$ stands for the jactivated unitof the hidden layer. In the forward-propagating transmission process, for the given input matrix ***X***, the activated unit of the hidden layer can be expressed as $a=sigmoid(WX+b)$. ***W***represents the weighted matrix between the input layer and the hidden layer, and the ***b*** represents the deviation matrix between the two layers. Then the average activation amount of the j unit in the hidden layer can be calculated as

$\rho_j = \frac{1}{n}\sum_{i=1}^{n}[a_j(x(i))]$ (1) Because most neurons are expected to be "inactive", we expect this average activation amount ρ_j to approach an approximate zero constant ρ, which is a sparse parameter at this point. To achieve such sparsity effect, additional penalty terms are added to the cost function of the encoder to penalize ρ_j deviations from parameter ρ. To this end, we chooseKullback–Leibler (KL) divergence [20] for the sake of punishment. The expression of the penalty term PNis listed as below $PN = \sum_{j=1}^{s_2} KL(\rho \| \rho_j)$ (2). In the formula, s_2 representsthe number of neure in the hidden layer; KL in divergence mathematics can be expressed as $KL(\rho \| \rho_j) = \rho \ln \frac{\rho}{\rho_j} + (1-\rho)\ln \frac{1-\rho}{1-\rho_j}$ (3). The penalty term is determined by the nature of the KL divergence: if $\rho_j = \rho$, then $KL(\rho \| \rho_j)$ =0. Otherwise, the KL divergence values will gradually increase as ρ_j deviates ρ. In the case of neural networks, the general cost function can be written as $C(W,b) = [\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\| h_{w,b}(x(t)-y(t)\|^2)\right)] + \frac{\gamma}{2}\sum_{l=1}^{m_{l-1}}\sum_{i=1}^{S_l}\sum_{j=1}^{S_{l+1}}(W_{ij}(l))$ (4). When the sparse penalty term is added, it can be expressed as $C_{sparse}(W,b) = C(W,b) + \beta\sum_{j=1}^{s_2} KL(\rho \| \rho_j)$ (5). In the formula, β stands for the weight of sparse penalty item. Throughout the encoding process, ultimately, we need get the optimized connection weight ***W*** and bias ***b***. And the sparse cost function C_{sparse} is a function with ***W*** and ***b*** as its parameters. Therefore, by minimizing the sparse cost function, we can get an optimal ***W*** and ***b***. In fact, it can be done by back propagation algorithm. At the same time, batch training method is used to update the weight in each iteration by gradient descent method. The updated equation is as follows:

$$W_{ij}(l) = W_{ij}(l) - \varepsilon \frac{\partial}{\partial W_{ij}(l)} C_{sparse}(W,b)$$

$b_i(l) = b_i(l) - \varepsilon \frac{\partial}{\partial b_i(l)} C_{sparse}(W,b)$ (6) In the formula, ε stands for the learning rate.

Therefore, the penalty term in the cost function is obtained by calculating the average activation quantity, and the sparse representation of the hidden layer can be obtained by optimizing the sparse cost function.

1.2 Restricted Boltzmann machine

Deep belief network is a neural network which is composed of a plurality of restricted Boltzmann machines (RBM) with multi hidden layers. Its core is to optimize the connection weights of the deep neural network by using the layer-by-layer greedy learning algorithm. Firstly, the unsupervised training layer-by-layer training is used to effectively excavate the fault diagnosis of the target equipment then, with the increase of the corresponding classifier, optimize the fault diagnosis ability of DBN through reverse supervised fine-tuning. Unsupervised layer-by-layer training can make direct mapping of data from input to output, and learn some nonlinear complex functions, which is the key to its powerful feature extraction capability.

Based on the literature [2] and [3], the RBM-related theory is briefly described as follows: one layer of RBM contains one visual layer 1 and one implicit layer h, as shown in Figure 2. Suppose layer 1 has r visible units, and the layer h has t hidden units. Then, the energy of a RBM can be expressed as

$$E(v,h|\theta) = -\sum_{i=1}^{r} a_i v_i - \sum_{j=1}^{t} b_j h_j - \sum_{i=1}^{r}\sum_{j=1}^{t} v_i W_{ij} h_j \text{.....(7)}$$

In the formula, v_i is the value of unit i in the visible layer; h_j is the value of unit j in the hidden layer and when the value is 1, it indicates that the unit is in active state, while 0 in an activated state; the parameters of RBM, W、a and b is referred to simply as θ. W is the connecting weight between the visual layer and the hidden layer; a is the bias vector of the visual layer; b is the bias vector of the hidden layer. Based on the RBM energy

representation, the joint probability distribution of (v, h) can be expressed as:

$$P(v,h\mid\theta)=\frac{e^{-E(v,h|\theta)}}{Z(\theta)}.....(8)$$

In the formula: $Z(\theta)=\sum_{v,h}e^{E(v,h|\theta)}$ isthe normalizing factor, namely, the partition function.

Then, the likelihood function of P(v|θ) can be expressed as:

$$P(v\mid\theta)=\frac{1}{Z(\theta)}\sum_{h}e^{E(v,h|\theta)}.....(9)$$

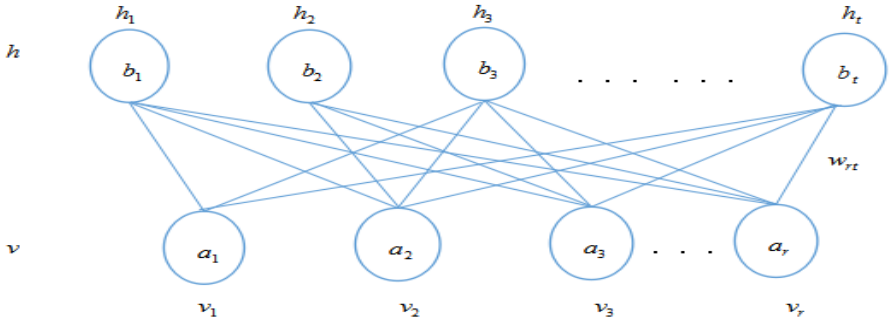


Fig. 2 RBM structure

Maximizing formula (9) via the gradient descent method and we can get parameter θ. To facilitate the calculation, correspondingly, we will maximize its log. The key step is the calculation of the partial derivative of $\ln P(v\mid\theta)$ concerning θ, i.e.,

$$\frac{\partial\ln P(v\mid\theta)}{\partial\theta}=\sum_{t=1}^{T}\left[\left\langle\frac{\partial(E(v,h\mid\theta))}{\partial\theta}\right\rangle_{P(h|v,\theta)}-\left\langle\frac{\partial(E(v,h\mid\theta))}{\partial\theta}\right\rangle_{P(v,h|\theta)}\right].....(10)$$

In the formula: $\langle.....\rangle_p$ is the mathematical expectation of distribution P; $P(v\mid h,\theta)$ represents the probability distribution of the hidden layer when the visual layer is defined as the training sample v;

$P(v\mid h,\theta)$ represents the joint probability distribution between the visual layer and the hidden layer.For convenience, "data" is used to refer to $P(v\mid h,\theta)$, and "model" is used to refer to $P(v\mid h,\theta)$. Now suppose there are only 1 sample, then the partial derivatives of $\ln P(v\mid\theta)$ about θ are respectively are:

$$\frac{\partial\ln P(v\mid\theta)}{\partial W_{ij}}=\langle v_i h_j\rangle_{data}-\langle v_i h_j\rangle_{model}.....(11)$$

$$\frac{\partial\ln P(v\mid\theta)}{\partial a_i}=\langle v_i\rangle_{data}-\langle v_i\rangle_{model}.....(12)$$

$$\frac{\partial\ln P(v\mid\theta)}{\partial b_j}=\langle h_j\rangle_{data}-\langle h_j\rangle_{model}.....(13)$$

In the formula: $\langle.\rangle_{data}$ represents the expectation of a data set; $\langle.\rangle_{model}$ represents the expected value defined in the model. In practical applications, it is difficult to obtain unbiased samples, so it is difficult to calculate $\langle.\rangle_{model}$. The CD algorithm is used to approximately sample the reconstructed data and update the network parameterθ。Take x_0 , a sample selected from a training sampleas an example. Its algorithm procedures are as below:

Step 1: Initialize the network parameter θ and set the initial value of the visual layer unit $v_0 = x_0$ and set the maximum number of training iterations for RBM.

Step 2: Calculate $P(h_{0j} = 1 | v_0) = \sigma(b_j + \sum_{i=1}^{n} v_{0i} W_{ij})$ in all the hidden layers. From the conditional distribution $P(h_{0j} | v_0)$ select $h_0 \sim P(h_0 | v_0)$. $\sigma(x)$ is a sigmoid function, the same below.

Step 3: Calculate $P(v_{1i} = 1 | h_0) = \sigma(a_i + \sum_{j=1}^{m} h_{0j} W_{ij})$ in all thevisual layers. From $P(v_{1i} | h_0)$, select $v_1 \sim P(v_1 | h_0)$.

Step 4：Calculate $P(v_{1i} = 1 | h_0) = \sigma(b_j + \sum_{j=1}^{m} v_{1i} W_{ij})$ in all thehidden layers.

Step 5: Update each parameter in accordance with the following formula:

$W \leftarrow W + \rho\left[P(h_0 = 1 | v_0) v_0^T - P(h_1 = 1 | v_1) v_1^T\right]$;

$a \leftarrow a + \rho(v_0 - v_1)$;

$b \leftarrow b + \rho\left[P(h_0 = 1 | v_0) - P(h_1 = 1 | v_1)\right]$。

Step 6: Repeat from step 2 to step 5 until the maximum iteration is reached or the reconstruction error is small enough. Finish the RBM training of the layer.

II. SYSTEM MODEL

This paper constructs a classification deep learning neural network (CDLNN) model. Its front part is composed of several layers of AE or RBM; the top layer is the last layer which represents the expectation output variable, namely the classification layer, whose framework is shown in Fig 3. Here, Softmax is selected as the classifier, which is suitable for multi classification problems, and can give the results of each classification in the form of probability. Combining with CDLNN, it will often get better discriminant performance.

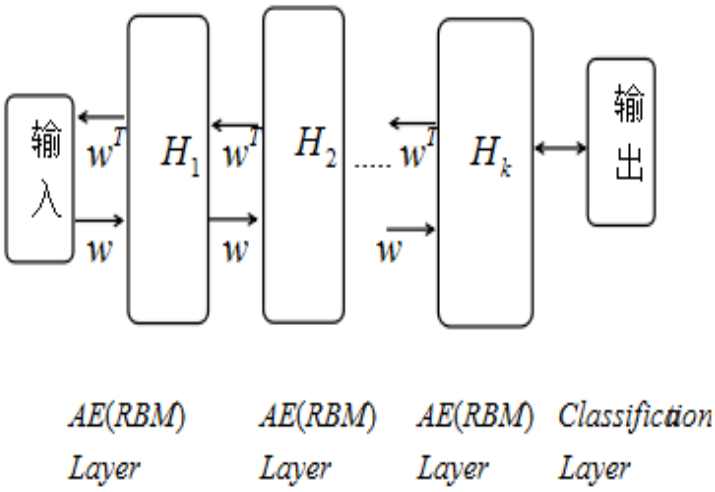


Fig.3 classification deep learning neural network

When CDLNN is used in multi classification problems, the training process can be divided into 2 stages: pre-training and tuning. Pre-training mainly uses unlabeled samples or label-free samples as the network input, and complete the parameter initialization of several layers of AE or RBM at the bottom through BP algorithm or CD algorithm; tuning is to fine tune the entire network parameters including the classification layer via the label samples, and make network discrimination achieve optimal performance.

Since 2006, when Hinton G.E et al proposed the deep belief network for the first time, they confirmed its powerful ability of feature extraction. Many applications during nearly ten years also confirmed this point. As a self-learning feature extraction algorithm, depth belief network has been widely used in many applications because of its powerful feature extraction ability and no need for participation of large amount of label data. The

process of DBN extracting fault features is shown in Fig 4.

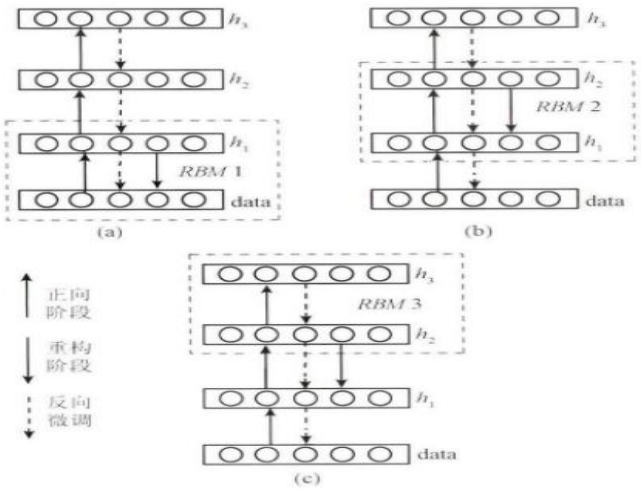
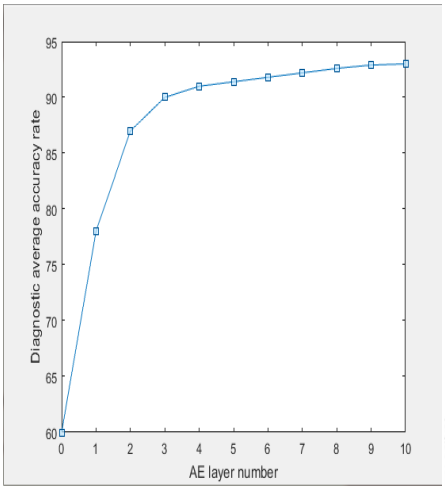


Fig.4 DBN feature extraction process layer by layer

First of all, make a layer-by-layer unsupervised training for DBN model; then, use the reverse fine-tuning algorithm to make a supervised training for the DBN model; finally, input the test data to the trained DBN model, and record the output vector in each hidden layer.

3. Analysis of experimental results



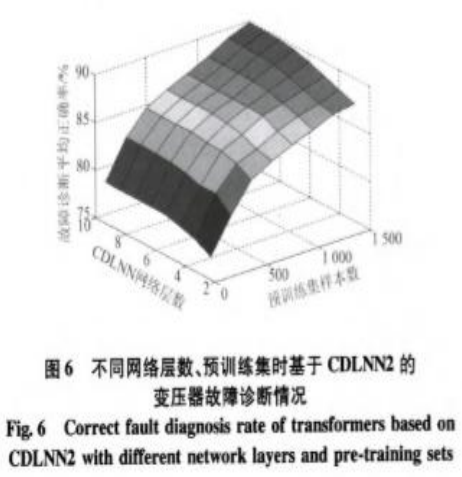


图 6 不同网络层数、预训练集时基于 CDLNN2 的变压器故障诊断情况

Fig. 6 Correct fault diagnosis rate of transformers based on CDLNN2 with different network layers and pre-training sets

Fig.5 Diagnostic average accuracy rate - AE layer number

(1) CDLNN l fault diagnosis in different layers of AE The average correct rate of power transformer fault diagnosis based on CDLNN l is tested when the number of AE layers is from 0 to 10, as shown in Figure 5. The "diagnostic average accuracy rate - AE layer number" curve shows that, when the number of AE layers reaches 3 layers, the average accuracy rate of fault diagnosis has been high, and then, as the number of AE layers increases, the accuracy rate increases slowly. In actual training, as the number of AE layers increases, the training time becomes longer. On the comprehensive consideration of the two factors, namely, fault diagnosis effect and training time, in the following test, 3 layers of AE are selected.

(2) As shown in Fig. 6, with the increase of the pre-training set, the least network layer that matches to maximum average accuracy rate of the CDLNN2 fault diagnosis changes from 6 layers to 8 layers, which shows the number of layers increases gradually. In the case of a given pre-training set, the average correct rate of fault diagnosis is increasing with the increase of network layer number. When the number of layers is reached a fixed value, the upward trend is slow.

III. CONCLUSIONS

(1) CDLNN model is constructed and its classification performance is analyzed. The typical dataset test shows that, CDLNN is applicable to multi classification problems.

(2) The new method for fault diagnosis of power transformer based on CDLNN uses a semi-supervised machine learning method, can effectively utilize the unlabeled samples obtained from the on-line monitoring of oil chromatogram for pre-training on the network and overcome the shortcoming that BPNN and SVM method cannot use unlabeled labels for the training. Therefore, it has a stronger learning ability and better fault diagnosis performance.

(3) The research findings show that, to the CDLNN diagnostic method, as the pre-training set increases, the average accuracy rate of fault diagnosis increases constantly. The method is applicable to the training of a large number samples and has a good scalability. Compared with the fault diagnosis methods of BPNN and SVM, its average diagnostic accuracy rate is higher, so that it can provide more accurate reference information for the overhaul of the power transformer.

REFERENCE

[1]. Yin Yujuan，Wang Mei，Zhang Jinjiang，et a1．An autonomic kernel optimization method to diagnose transformer faults by multi—kernel learning suppoa vector classifier based on binary particle swamioptimization[J]．Power System Technology，2012，36(7)：249-254．

[2]. Yu Bingjie，Zhu Yongli．Transformer fault diagnosis using weighted extreme learning machine[J]．Computer Engineering and Design，2013，34(12)：43404344．

[3]. Hinton G E，Salakhutdinov R．Reducing the dimension—ality of data with neural networks[J]．Science，2006，313(5786)：504-507．

[4]. SHAO H，JIANG H，ZHANG X，et al．Rolling bearing fault diagnosis using an optimization deep belief network［J］．Measurement Science & Technology，2015，26 (11) : 115002115018．

[5]. LI C，SANCHEZ R V，ZURITA G，et al．Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis［J］.Neurocomputing，2015，168(C) :119127．

[6]. BENGIO Y．Learning deep architectures for AI[J].Foundations &Trends in Machine Learning，2009，2(1) :121127．

[7]. HINTON G E．A Practical Guide to Training Restricted Boltzmann Machines［M］.Neural Networks: Tricks of the Trade，2012:599

Liu Bingyao1. “Power Transformer Fault Diagnosis based on Deep Learning.” International Journal of Research in Engineering and Science (IJRES), vol. 05, no. 08, 2017, pp. 42–48.