Researches Advanced in Image Recognition based on Deep Learning

Zheke Yi *

The University of Sydney, New South Wales, Australia

* Corresponding author email: zhyi3713@uni.sydney.edu.au

Abstract. In the discipline of computer vision, which tries to create a model to predict the category of objects present in a given image, image recognition has always been a research hotspot. Early image recognition models were mainly based on manual features, and their recognition accuracy and generalization ability often fluctuated greatly with changes in the scene, which could not meet the actual application requirements. Convolutional neural networks have progressed quickly, which has accelerated the development of deep learning-based image recognition. in this paper, we detail the development of image recognition technology. Specifically, we introduce the three approaches of convolutional neural network, recurrent neural network, and graph neural network as the traditional methods of image recognition. including their design ideas, key steps, advantages and disadvantages. We also compare the recognition accuracy of different methods on the ImageNe dataset to try to explore the application boundaries of different methods. Finally, we go over the difficulties in the area of image identification and project its future growth.

Keywords: Image Recognition; Deep Learning; Computer Vision.

1. Introduction

One of the most significant research goals in the field of computer vision has always been the development of image recognition technology, which enables computers to process, analyze, and comprehend images in order to distinguish targets and objects in various patterns. Image recognition is now widely employed in a variety of industries, including biomedicine, face identification, driverless cars, etc. Additionally, picture recognition is useful for a variety of computer vision tasks, including object detection, segmentation, tracking, and others.

For the traditional image recognition, there are three mainstream processing methods: the statistical methods, the syntax recognition methods, and the geometric transformation methods. Image recognition in the statistical method takes mathematical decision-making theory as the basis, which first digitizes the image and begins to establish a statistical recognition model to perform statistical analysis on the image. The obtained corresponding image features then will be used to classification. Since this model can handle small-scale multiple classification tasks, it is not sensitive to missing data, and the algorithm is relatively simple. But when the image is complex and there are many types, it is difficult to extract image features, so it is difficult to realize image classification. So, it is difficult to identify images featuring fingerprints or chromosomes. The syntax recognition method decomposes complex images into single-layer or multi-layer simple sub-images, highlighting the structural relationship with space. This kind of methods can not only classify and recognize images, but also can be used to recognize object structure, scenery and objects. However, for larger images, such as noise and interference, these special factors will affect the sub-image in the syntax recognition process, so that the probability of misjudgment becomes higher and the accuracy of classification recognition becomes lower. The geometric transformation method is that when the angles of the camera and the shooting object are not at the same angle, the resulting image will be geometrically distorted. Therefore, it is necessary to change the image by performing geometric transformations such as translation, rotation, and scaling. At the same time, this method is also called the Hough transform method, which is a typical method in the geometric transformation method, and this method is a classic in the geometric transformation method. These methods can quickly match the shape of the graph, and also has strong anti-interference properties, and is not overly sensitive to

noise, imperfect straight lines, or other non-linear image structures. However, due to the algorithm's high time and space complexity, only the straight line's direction can be determined, and the length information of the line segment is susceptible to loss.

With the advancement of technology, compared with these traditional algorithms, today's deep learning has more advantages than the previous traditional algorithms, such as: 1) convolutional neural network: features can be automatically selected, when a large amount of learning is required, it has strong stability. 2) recurrent neural network: parameters in different situations can be shared when extracting nonlinear features. 3) Graph Convolutional Neural Network: it is easy to handle tasks such as node classification and edge prediction, and can be used on nodes with any flutter structure. Focusing on the above three types of representative frameworks, a large number of image recognition methods based on deep learning have been proposed, which have significantly improved the recognition accuracy and speed.

In this paper, we will introduce the research progress of image recognition based on deep learning in detail. Specifically, we will introduce representative image recognition algorithms in Section 2, including their design ideas, network structures and key steps, etc. In Section 3, we compare the experimental results of these algorithms on different experimental datasets to analyze their strengths and weaknesses. Finally, in Section 4, we summarize the remaining research issues in the field of image recognition research and discuss its possible future directions.

2. Methods

2.1 Recognition based on CNN

The convolutional neural network and the conventional neural network share many characteristics, including the threshold, forward propagation, residual calculation, and back propagation, up until the residual converges and satisfies the accuracy requirements. Convolution, pooling, and fully connected layers make up a convolutional neural network's basic architecture [1].

A crucial component of the calculation using the convolutional layer is the convolution kernel. The initial image is transformed into a hyperplane coordinate system by the convolutional layer's operation. The hyperplane can concentrate similar images to the greatest extent. Convolutional layers can effectively extract feature images. The convolutional layer is the core layer of the convolutional neural network, which contains a large number of calculations. At the same time, the parameters of the convolutional layer include the convolution kernel, step size and padding. At the same time, the convolutional layer has two major characteristics, namely weight sharing and local linking [2-3]. Because each planar layer contains a convolution kernel, the extracted features are mainly obtained by the budget in the convolution layer. Due to the characteristics of the convolution kernel, a large number is not necessarily good. Sometimes the number is too large, which will lead to an increase in the difficulty of network training. Therefore, the number of convolution kernels should consider the actual situation at that time.

The downsampling layer is another name for the pooling layer. Utilizing the convolutional layer's processed data as input, it performs a pooling operation to partially compress the findings. By doing so, the spatial size of the data is reduced, and fewer parameters are needed, increasing computing efficiency and effectively reducing overfitting. The pooling layer is used to perform network training on a large number of pictures in the convolutional neural network. At the same time, in order to reduce the burden on the network, it performs dimensionality reduction processing and maintains the original features. The efficiency of the network is greatly improved by pooling the pictures. At the same time, pooling is also divided into different methods: maximum pooling, mean pooling [4-6]. The fully connected layer at the end of the convolutional neural network will turn the two-dimensional feature map output by the convolution into a one-dimensional vector to improve the quality of feature extraction and make it simpler to transfer the input to the final classifier or regression. Each convolution kernel performs a convolution computation throughout the forward propagation process

to produce a 2-dimensional feature map. These convolution result maps are layered to produce the output following a specific network training.

The methods that recognizing images using the convolutional neural networks was first proposed in 1962. Wiesel et al. studied the brain vision in cats and sorted out the electrical activity of each neuron in the cat's brain. Inspired by this interesting phenomenon, the idea of receptive field was proposed. In 1980, a neural network structure was proposed by Fukushima et al. This structure includes convolutional layers and pooling layers, thus named the neurocognitive machine algorithm. Lecun et al. first proposed the idea of "convolution" in 1998 when they proposed the LeNet-5 network, which is composed of two fully linked layers and two convolutional layers. Handwritten digit recognition is addressed using LeNet-5[7]. All of the input photos are single-channel, 2828 grayscale pictures. The LeNet-5 model outputs 10 probabilities through the output fully connected layer, which corresponds to the predicted probability from 0 to 9, through a total of 7 layers (excluding the input layer), including two convolutional layers, two down-sampling (pooling), and three fully connected. The network structure of it is shown in Fig. 1. The input image is 3232 pixels in size, the convolution window is 55 pixels in size, the convolution kernel is translated on a two-dimensional plane, and each component of the convolution kernel is related to the corresponding positions of the convolution image and is multiplied and then summed. Each layer has training parameters. An image made completely of the outcomes of the product summation of the convolution kernel at each place can be produced by continuously moving the convolution kernel. It convolves the input image using a sliding convolution window approach after pooling and fully linked layers. LeNet-5 produced an accuracy rating of about 99.2% on the MNIST dataset. This method is mainly used in object detection, face recognition, target detection, speech recognition and other fields [8].

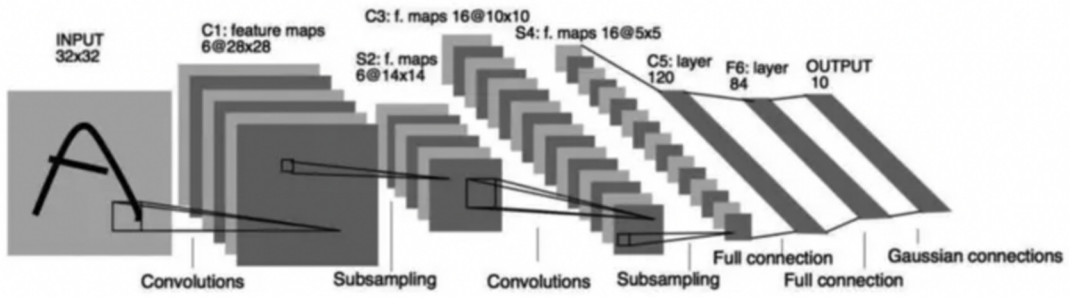


Fig 1. The network structure of LeNet-5

With an absolute margin of 10.9 percentage points over the runner-up, AlexNet won the 2012 ImageNet competition. Convolutional neural networks and deep learning have gained popularity since that time. Compared to LeNet, AlexNet designed a deeper network. AlexNet is aimed at the classification problem of 1000 categories. The input image is specified as a 256×256 three-channel color image. In order to enhance the generalization ability of the model and avoid over-fitting, the author uses the idea of random cropping for the original 256×256 image. Perform random cropping to obtain an image with a size of $3 \times 224 \times 224$, which is input to the network for training. AlexNet has 5 convolutional layers and 3 fully connected layers, interspersed with pooling operations. In addition to convolution, pooling, and full connection operations, this model also introduces the ReLU activation function and local response normalization (LRN).

In the next year, ZFNet was proposed by Zeiler [9-10], which allows CNN to be visualized, uses smaller convolution kernels, and retains more features, which improves CNN's hierarchical abstract learning ability. Since then, VGG-Nets has been proposed and is the basic network in the 2014 ImageNet competition for the first place in the localization task and the second place in the classification task. VGG can be regarded as a deepened version of AlexNet. Both are conv layer+FC layer, because the number of layers is as high as ten layers, it seemed that this was a very deep network at the time. Figure 5 describes the VGG network structure and birth process. The pre-training method used by classic neural networks is frequently used by VGG to address issues like initialization (weight

initialization). In order to gradually deepen on this basis, it is best to train a small portion of the network first, then check to see if it is stable. The network in stage D is VGG-16, and Figure 5 depicts this process from left to right. When the network is in stage D, the effect is optimal. VGG-19 is the network that was obtained at the E stage. The 16 of VGG-16 denotes the total number of conv+fc layers, not including the number of layers for max pool, which is 16.

Unlike AlexNet and VGG-Nets, which rely solely on deepening the number of layers of the network structure to improve network performance, GoogleNet introduces the Inception structure to replace the traditional operations of simple convolution and pooling while deepening the network (22 layers). Multiple convolution kernels are used to extract the information from the image's various scales, which is then combined to produce a better representation of the image. Based on the concept of Network in Network, GoogLeNet enhances the convolution kernel and converts the initial linear convolution layer into a multilayer perceptron, giving the convolution kernel stronger feature extraction capabilities. Additionally, GoogLeNet replaces the final fully connected layer with the global average pooling layer (Average pool), greatly reducing parameters and reducing overfitting.

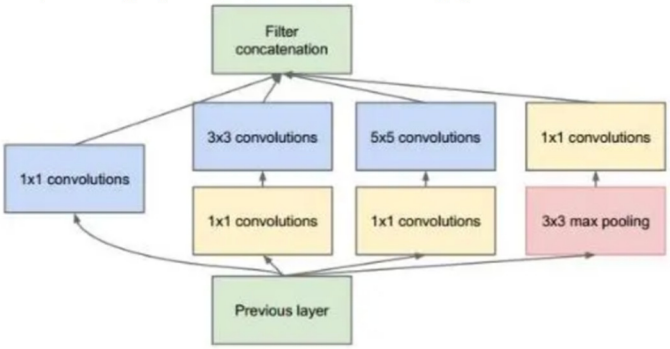


Fig 2. The network structure of Inception module

The accuracy of the network should, in theory, rise in tandem with the network's depth. The gradient of the earlier layers will, however, be minimal since the gradient is propagated from the rear to the front when the depth of the network is increased. This indicates that the vanishing gradient problem results from a learning halt in these levels. Simply increasing the network depth will result in higher training mistakes because, in addition, as the network becomes deeper, the parameter space grows and the optimization issue becomes more complex. In 2015, He et al. proposed ResNet, which introduced residual units to solve the degradation problem. As shown in Fig.3, The data has traveled down two different paths: a standard path and a shortcut (shortcut), a direct link path that implements unit mapping directly. If we assume that an individual network module's input and output are y=H(x), then using the gradient approach to directly calculate H(x) will result in the degradation issue discussed above. If this shortcut structure is utilized, the variable parameter portion's optimization target is no longer H(x). Instead, if F(x) is used to represent the part that has to be optimized, H(x)=F(x)+x, or F(x)=H(x)-x. Assuming unit mapping, y=x is comparable to the observed value, hence F(x) corresponds to the residual, hence the name residual network.

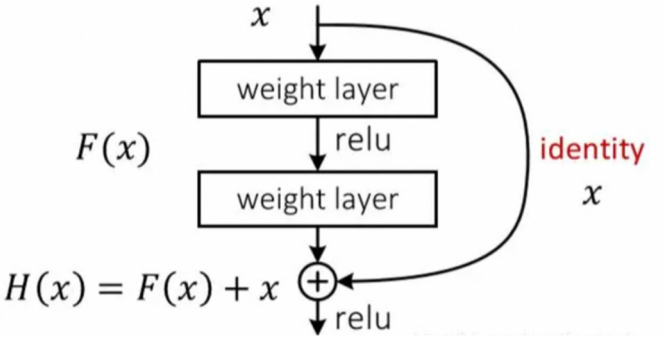


Fig 3. The network structure of residual module

Since Resnet was proposed, variant networks of ResNet have emerged one after another, which has greatly promoted the accuracy and speed of image recognition. Representative works include Dense Convolutional Network (DenseNet) and Squeeze-and-Excitation Networks (SENet), etc. Take DenseNet as an example, it has numerous connections and is a convolutional neural network. The input of every layer in this network is the union of the outputs of all preceding levels, and the feature map learned by this layer will also be directly transmitted to any other layer in the network. The input for each layer after that is utilised.

2.2 Recognition based on RNN

The recurrent neural network (RNN) is applied to the scene when the input data is patient and sequential, that is, the previous input and the next input are related. Unlike other structures, the hidden layer of RNN is cyclic. When training this model, the issue of gradient explosion and gradient disappearance is prone to arise. As a result, the problem of low transferability occurs during training, and the gradient cannot be transferred in longer sequences, which makes RNN unable to detect the impact of long sequences. The vanishing gradient problem is harder to detect than exploding gradients and can be solved by RNNs of other structures.

Because of the problem of gradient disappearance, RNN may only have short-term memory, and there will be a problem of "long-term dependence". At this time, LSTM has been improved on the basis of RNN. LSTM uses three "gate" structures—the "input gate," "output gate," and "forget gate"—to regulate the state and output at various moments, in contrast to the recurrent layer in the fundamental structure of RNN. The "forget gate" regulates how many pieces of information can be transferred to the unit state at the present instant from the previous moment, while the "input gate" and "output gate" regulate how many pieces of information can be saved to the unit state at the present time. determines the unit state's output number of messages to the output value for the current state.

In 2018, LSTM was improved and promoted by ALEX Graves, an autoencoder framework that can be used to detect whether a computer network has been invaded. Later, Koutnik et al. proposed a simple and effective modification to the standard RNN framework, namely ClockworkRNN (CW-RNN), which can effectively solve the problem of gradient disappearance. Then Zhao et al. made a series of improvements to this, and proposed GRU. GRU can simplify the structure of LSTM and maintain the same structure as LSTM. This method simplifies two gates, namely "update gate" and " reset door".

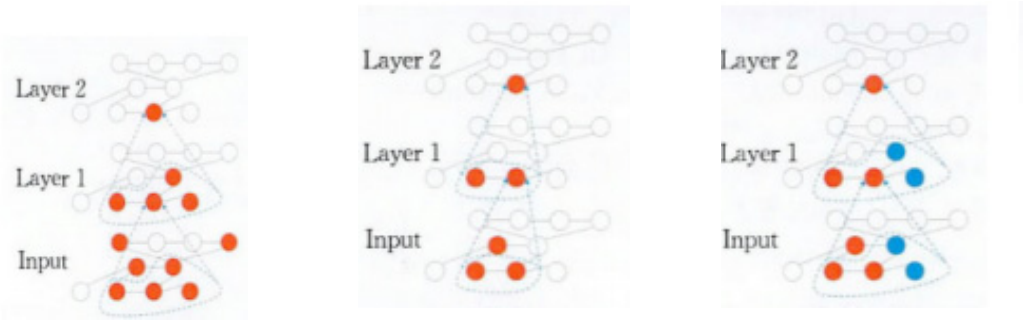
2.3 Recognition based on GCN

A convolution operator plus a pooling operator makes up the graph convolutional neural network (GCN). The convolution operator's job is to characterize the node's local structure. The pooling operator's primary duties are to learn the network's hierarchical representation and minimize the parameters. Utilizing the graph theory-based convolution theorem, Bruna et al. created the first convolutional neural network in 2013. High space-time complexity is a drawback of the original spectrum approach. In order to simplify space-time computation, ChebNet and GCN parameterize the convolution kernel.

At present, the mainstream methods of graph convolution neural networks are divided into spectral methods and spatial methods. The spectral technique uses the convolution theorem on the graph to define convolution in the spectral domain. By specifying an aggregation function, the spatial method aggregates each terminal by starting from the node domain and moving outward to include the point and its neighbors. The map convolution is defined from the spectral domain using the graph convolutional network spectral approach.

When the input signal is applied to the convolutional neural network's convolution kernel in spectral space, and uses the convolution theorem to achieve the purpose of image-selection convolution to complete the information aggregation between nodes. The purpose of modeling graph convolution at the beginning is to describe the information set of adjacent nodes through the graph structure. Locality cannot be satisfied because of the general convolutional neural network's features.

Smoothness was suggested as a method by Henaff et al. [10-12] to limit the interpolation convolution. By using a kernel, this technique allows for the localization of graph convolutional neural networks while reducing the number of parameters. The chart convolutional neural network space approach creates an aggregation function for each core node and any nearby nodes before starting from the node domain and defining graph convolution in the spectral domain.



(a) Accurate neighbors (b) Neighbors (c) Control variance

Fig 4. The difference between CV, NS, and GCN

GraphSAGE [13] uses the method of using batches and adopts the Neighbor Sampling (NS) method to keep the number of nodes that need to be calculated each time within a certain range. The number decreased. Figure 4 shows the difference between CV, NS, and GCN. The blue nodes in the figure represent the expression of history, and the red represents the current status of modern research. Among them, the CV method is the expression of the neighbor nodes that need to be sampled, and the NS is to estimate the expression of the entire neighbor by sampling the neighbor points, but this method has a large variance. The hierarchical sampling method is not to sample in units of nodes, but to sample through the nodes required between each layer, that is to say, a sub-graph is sampled from the original image from each layer, and only by sampling to the node to operate. In this case, the number of nodes sampled and the number of network layers becomes linear, and the calculation is also reduced.

3. Experiments

3.1 Common Datasets

The most typical image recognition datasets are ImageNet and PASCAL datasets. After years of hard work, these massive and freely available datasets contain millions of images, each tagged with keywords relevant to the image content.

(1) ImageNet: Created by researchers at Princeton University in 2009, this visualization dataset has more than 14 million URL images collected from search engines such as Flickr. During the dataset creation process, staff and volunteers annotated the submitted images in detail and classified them into about 1000 object classes.

(2) PASCAL: PASCAL was jointly created by various universities in EU countries. Compared with the ImageNet dataset, PASCAL pales in comparison - only 20 object classes, a total of 20,000 training images.

(3) Mnist: The width and height of each sample image in the handwritten digital database MNIST are also 28 by 28. It contains 10,000 test sample sets and 60,000 training sample sets.

3.2 Performance Comparison

In this section, we compare the experimental results of different representative methods on the most challenging ImageNet dataset, as shown in Figure 5. In early stage, the state-of-the-art model is the AlexNet and MSRA, where the Top1 Accuracy is low than 75%, far away from the human recognition accuracy. With the development of technology, the recognition accuracy has increased to

more than 90% in the latest work. The gain of performance mainly comes from the high-quality feature representation provided by more powerful convolutional neural networks.

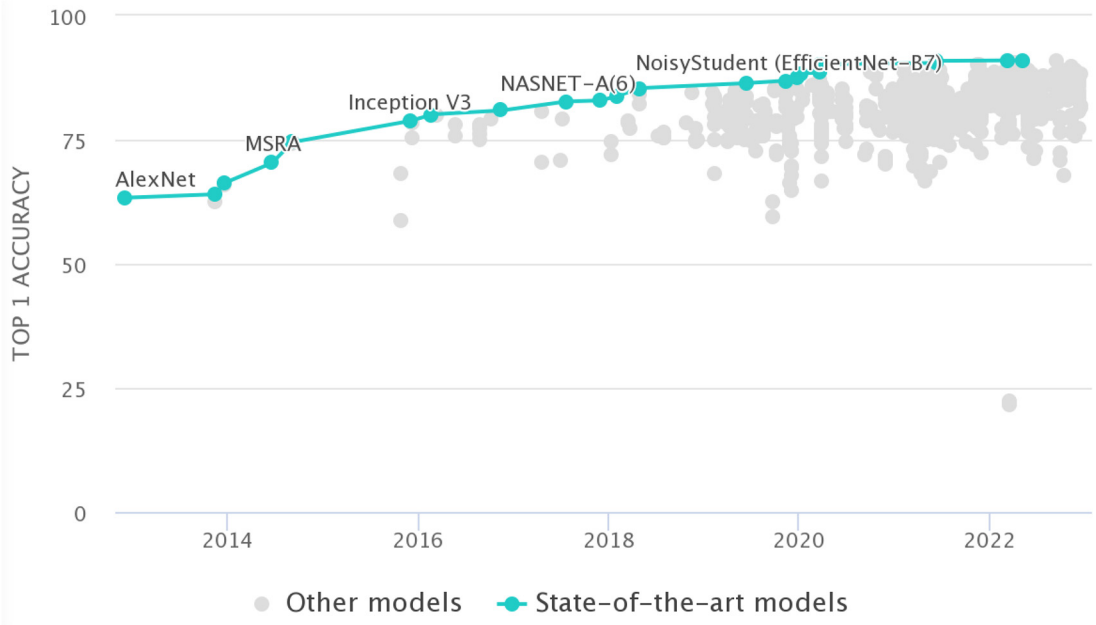


Fig 5. Comparison of the recognition performance on the ImageNet

4. Discussion

Image recognition plays a very important role in many fields. With the development of artificial intelligence, convolutional networks, cyclic neural networks, and graph convolutional networks have irreplaceable advantages. It is necessary to increase the independent learning ability of the network to make the network easier and effectively solve practical problems. With the gradual improvement of CNN's technology for target detection, it is rapidly developing in many different fields such as unmanned driving, education, medical care, and intelligent facilities. At the same time, deep learning is an important branch in the field of machine learning. Through machine learning of these classic network structures, computers can quickly learn their specific characteristics and laws from a large amount of data.

At present, the main challenges in the field of image recognition include: (1) Angle of view change: the camera can display the same object from multiple angles. (2) Size change: The visible size of an object usually changes (not only in pictures, but also in the real world). (3) Deformation: Many things' shapes are dynamic and subject to significant modification. (4) Occlusion: The target object can be blocked from view. Sometimes, only a little portion of the object—as small as a few pixels—is discernible. (5) Lighting conditions: At the pixel level, the influence of lighting is very large. (6) Background interference: It may be difficult to distinguish items when they are blended into the background. (7) Intra-class variations: Individual members of a class of things, like chairs, have a wide range of shapes. This category includes a wide variety of things, each with a distinctive shape.

5. Conclusion

This article gives a quick overview of three different neural network structures, their concepts, their origins, the individuals who came up with these ideas, as well as their benefits and drawbacks. The convolutional neural network is susceptible to several issues during the training process because of the high number of parameters. The process of feature aggregation in the creation of deep networks in the graph convolutional network is simple to generate a significant amount of calculation, which has to be enhanced by EGCN. When it comes to tackling challenges involving sequence input,

recurrent neural networks shine. By modifying the RNN's parameters and the recurrent layer's structure, various issues can be resolved in the future.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.

[2] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[4] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[5] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[6] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[7] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[8] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[9] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[10] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[11] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[12] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[13] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.