An Unsupervised feature extraction method based on self coded neural network

Long Hao, F Jiang
College of Electronic Engineering, Guangxi Normal University, Guilin 541004, China
fjiang@gxnu.edu.cn

Abstract. The way to tap the inherent law of ECG data and feature extraction is of vital in the absence of prior knowledge of the situation. This paper presents a self coded neural network model, which is reconstructed after compression of ECG data. The compression process eliminates the redundancy of data within the original data firstly from the low dimensional with more concise representations. In order to verify the features extracted from the neural network, we select the simplest and most direct distance based K nearest neighbour classification method. By using a self compiled neural network to extract features from the data set, the classification accuracy of K nearest neighbour classification can be greatly improved. The experimental results show that the unsupervised feature extraction method based on self coded neural network can effectively extract the features of the data in practical applications.

1. Introduction
This paper introduces a self encoded neural network model for reconstruction of the first treatment after compression of ECG data. The unsupervised learning feature of the neural network is used to provide a new method for unsupervised feature extraction in the diagnosis of cardiovascular diseases.

In Section II the self compiled neural network model is provided. And in Section III the unsupervised feature extraction method based on self coded neural network is provided [7,8]. By using a self compiled neural network to extract features from the data set, the classification accuracy of K nearest neighbor classification can be greatly improved. The experimental results show that the unsupervised feature extraction method based on self coded neural network can effectively extract the features of the data in practical applications. The detail processes are illustrated in Section IV.

2. Self compiled neural network model
The self encoded neural network is very similar to the BP neural network, and the only difference is that in the training and learning stages of the network, the self coded neural network specifies the expected output of the network is equal to the input of the network. The goal of the self coded neural network learning is to make the output of the network equal to the input of the network. In this way, the training of self coding neural network does not require training data to provide the desired network output. Therefore, the training of the neural network belongs to the category of the unsupervised learning.

Fig. 1 is a schematic diagram of a self coded neural network. As the output of the network is equal to the input requirements of learning rules, the number of neural units of input layer autoencoder neural networks must be equal to the number of neural cells and the output layer, and the number of neurons in the hidden layer can be arbitrary.
3. Unsupervised feature extraction method based on self coded neural network

3.1 Training of the self coding neural network model
For the self coded neural network in Figure 1, the output \( h_{W,b}(x) \) of the neural network can be calculated according to the forward algorithm when the model parameter \((W,b)\) is input to the network.

\[
d^{(3)} = f\left(W_3^{(3)}x_1 + W_3^{(2)}x_2 + W_3^{(1)}x_3 + W_3^{(0)}x_4 + b^{(3)}_3\right)
\]
\[
\hat{x}_i = f\left(W_i^{(2)}a_1 + W_i^{(1)}a_2 + W_i^{(0)}a_3 + b_i^{(2)}\right)
\]
\[
h_{W,b}(x) = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{x}_5, \hat{x}_6)
\]

First of all, the use of the original input to train the first layer stack autoencoder neural network, the training method of each layer is an unsupervised learning algorithm since the encoding of neural network after training can be the network parameters \((W^{(C)}, b^{(C)})\) of the layer.

Secondly, the output value of the intermediate hidden layer neuron is inputted into the second layer of the network to train the layer. After the training, the network parameters \((W^{(2)}, W^{(1)}, b^{(1)})\) of second layers can be obtained.

Third, the same training strategy is used to complete subsequent training. After each layer of network training, the network parameters of the layer are recorded until all the layers are trained, so that the model parameters of the whole network can be obtained.

3.2 Unsupervised feature extraction based on self coded neural network
In fact, the human body is a very complicated system, when the body produces cardiovascular lesions of various signs and ECG parameters have a high correlation, and since self encoded neural network can be found and the reduction of ECG using the correlation data internally to extract the features of ECG data so it is feasible to apply feature extraction from encoding neural network is applied to ECG data[3].

The feature extracted from the neural network based on the self encoded neural network is the output of the network hidden layer.
Fig. 2 and 3 are schematics diagrams of a self encoded neural network for extracting features.

The original data is \( x = (x_1, x_2, x_3, x_4, x_5, x_6) \), where the number of neurons in the hidden layer of the network is 3, so that the vector \( a = (a_1, a_2, a_3) \) is composed of the activation values of the 3 neurons in the hidden layer as the characteristic representation of the original data.

From figure 2 and 3, we can see the process of feature extraction in autoencoder neural networks can be described as: the first training based on autoencoder neural networks, then remove the last layer of network, which features between the hidden layer output is extracted.

The dimension of the output vector of the hidden layer is larger than the dimension of the input data. It is no longer a representation of the input data, nor can it be used as a feature representation of the input data.

Since the training of neural network for self coding belongs to the unsupervised learning, the feature extraction method based on self coded neural network belongs to unsupervised feature extraction method.

### 3.3 Method for selecting number of hidden layer neurons in network

In this paper, we propose a fast method to determine the number of neurons in the network hidden layer, which is called "the method of doubling the value".

The method firstly needs to determine the range of the number of neurons in the network hidden layer, that is, its maximum value and minimum value[4]. The minimum number of neurons in the hidden layer generally defaults to 1, and the maximum is equal to the number of neurons in the network input layer.

Determine the value of the range, to find the midpoint of the range of values \( m_1 \), the point will be divided into two ranges of values \( [n_{\text{min}}, m_1] \) and \( [m_1, n_{\text{max}}] \), and then find the midpoint \( m_2 \) of these two ranges. The above 5 values are considered as hidden layer units. The features of their corresponding neural networks are classified by K nearest neighbor method, by comparing the results of the 5 values, we can judge that the number of hidden layer units is in the interval \( [n_{\text{min}}, m_1] \) or the value \( [m_1, n_{\text{max}}] \) of the self encoded neural network is better than that of the extracted neural network. Because the number of the hidden layer neurons is chosen as the number of hidden layer neurons in the midpoint of the chosen value, the number of attempts is greatly reduced, so the determination of the number of neurons in the hidden layer is accelerated.

### 4. UCI data Set validation

#### 4.1 Validation of feature extraction methods

We choose the simplest and most direct classification method to emphasize the effectiveness of feature extraction from neural networks, which the distance bases on K nearest neighbor classification.

The classification performance is poor, when it bases on this method. If the extracted features can significantly improve the classification accuracy of the K nearest neighbor classification method, then the effectiveness of the feature extraction method proposed in this chapter can be proved.

In this experiment, we select 3 data sets from the standard data set UCI to test the K nearest neighbor classification. These 3 data sets are wine, Iris, CBF. In the experiment, K nearest neighbor classification method was used to classify the above 3 data sets, and the correct rate of each data set was recorded. Then 3 feature sets are extracted by using a self coded neural network, and then the K nearest neighbor classification method is used to classify these features, and the correct rate is recorded.

In the classification process with the K nearest neighbor method, when the parameter value of K is \( K=19 \), each classification test cycle for 150 times at the same time in order to make the classification results stable. Table 1 shows the classification results of the test set and the two classification experiments.

| Table 1. Comparison of the results of the two classifications. |
The experimental results show that the unsupervised feature extraction method based on self coded neural network can effectively extract the features of the data in practical applications.

4.2 Validation of the number of hidden layer neurons

In section 4 of the K nearest neighbor classification experiment, can find the correct rate of K neighbor classification feature extraction self coding based on neural network classification in fixed cycle 150 times the test is not, but in a small range. Moreover, when the number of neural units in the hidden layer of the neural network is changed, the classification accuracy will be greatly changed[5].

The Wine data set is used as the sample data, the number of neurons in different hidden layers is selected, and the change of the accuracy of the K nearest neighbor classification based on the extracted feature of the neural network is recorded. Preliminary validation of the proposed "half value" method is effective, how to select the number of hidden units in order to provide some experience for the practical operation data of application behind the autoencoder neural networks of machinery.

Since the data contains 13 eigenvalues, the number of neurons in the input layer of the network must be 13. Since the number of neurons in the hidden layer must be no more than the number of neurons in the input layer, the number of neurons in the hidden layer of the network is likely to be the value $m = 1, 2, \ldots, 13$ of the number of neurons in the hidden layer.

When the 13 values are taken respectively, the fluctuation rate of the correct rate of the K nearest neighbor classification is shown in Figure 4-18.

| Data set      | Direct classification | Self encoder |
|---------------|------------------------|--------------|
| Wine          | 0.67                   | 0.92         |
| Synthetic Control | 0.82               | 0.92         |
| Iris          | 0.79                   | 0.98         |
As can be seen from above figures, the effectiveness of appropriately increasing the number of hidden layer neurons can make the feature extraction of hidden layer enhancement.

The next step is that the number of hidden layer neurons is set to 9 or 10 when the feature of wine data is extracted by using a self encoded neural network, so that the minimum dimension of the original data can be effectively represented.

Table 2. The First K nearest neighbour classification results

| Hidden layer element number | 1   | 4   | 7   | 10  | 13  |
|-----------------------------|-----|-----|-----|-----|-----|
| Classification accuracy     | 0.59| 0.80| 0.88| 0.90| 0.90|

According to the “folded values” method, the next step should be to find the appropriate number of hidden layer units in the interval [7,12]. The intermediate points of the interval are chosen as the number of hidden layer units and their K nearest neighbour classification results are shown in Table 3.

Table 3. The Second K nearest neighbor classification results
### Table: Classification Accuracy

| Hidden layer element number | 7   | 9   | 10  | 12  | 13  |
|-----------------------------|-----|-----|-----|-----|-----|
| Classification accuracy     | 0.88| 0.90| 0.90| 0.90| 0.90|

We can see the "half value" method can quickly determine the number of hidden layer neuron network.

### 5. Conclusion

The method of feature extraction based on self coded neural network is studied in view of the lack of prior knowledge of signal in feature extraction. To achieve the goal of unsupervised feature extraction in the equipment condition data without prior knowledge, we use unsupervised learning characteristics auto encoder neural networks and put forward the "half value method" to achieve rapid determination of self encoding neural network hidden layer unit number. Finally, the effectiveness of the feature extraction method and the hidden layer unit number selection method is verified by standard data sets.

### Acknowledgments

Firstly, I would like to thank Dr. Jiang for his constant help and efforts in improving the work and the paper writing process, Dr. Jiang gave me a lot of encouragement. Secondly, I thank the anonymous reviewers for their valuable comments in improving the final quality of this paper. This work is supported by the Guangxi 100 Youth Talent Program (F-KA16016) and the Colleges and Universities Key Laboratory of Intelligent Integrated Automation, Guilin University of Electronic Technology, China (GXZDSY2016-03), National Natural Science Foundation of China (61762018), the research fund of Guangxi Key Lab of Multi-source Information Mining & Security (18-A-02-02).

### References

[1] Hua X, The value of combined detection of blood lipid markers in the prediction of cardiovascular and cerebrovascular diseases [J]. Chinese medical journal, 2012, 10 (13): 231-232.

[2] Zhu L. M, Wang X.Y, Pulse pressure and risk factors of cardiovascular and cerebrovascular diseases [J]. Chinese Journal of cardiovascular diseases, 2001, 29 (9): 573-574.

[3] Hand D J, Till R J, A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems[J]. Machine Learning, 2001, 45(2):171-186.

[4] Chae C U, Pfeffer M A, Glynn R J, et al. Increased pulse pressure and risk of heart failure in the elderly.[J]. Jama, 1999, 281(7):634-643.

[5] Zhu C, Byrd R H, Lu P, Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization[J]. ACM Transactions on Mathematical Software1997, 23(4): 550-560.

[6] [Online] Available: http://archive.ics.uci.edu/ml/datasets/Wine [Accessed: 10- Oct- 2018].

[7] Sermanet P, Kavukcuoglu K, Chintala S, et al. Pedestrian detection with unsupervised multi-stage feature learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3626-3633.

[8] Chen D, Manning C, A fast and accurate dependency parser using neural networks[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 740-750.

[9] Kong X, Hu C, Duan Z, Singular Feature Extraction and Its Neural Networks[M]//Principal Component Analysis Networks and Algorithms. Springer Singapore, 2017: 281-323.

[10] Abedin M Z, Dhar P, Deb K, Traffic sign recognition using hybrid features descriptor and artificial neural network classifier[C] 19th International Conference on Computer and Information Technology (ICCIT), 2016: 457-462.