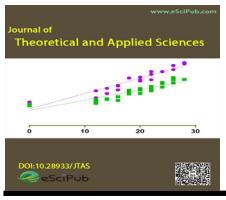
Research Article **JTAS (2019) 2:10**



Journal of Theoretical and Applied Sciences
(ISSN:2637-692X)

Abnormal data cleaning in thermal power plant based on self-organizing maps

Song Yu[1], Guozhu Chen[2]*, Zhou Bao[1], Baogang Song[2]

1.College of Computer and Information Technology, China Three Gorges University,Yichang, 443002, China. 2.College of Hydraulic & Environmental Engineering, China Three Gorges University, Yichang, 443002, China.

ABSTRACT

This paper constructs a self-organizing maps (SOM) neural network model for the anomaly data cleaning in thermal power plant detection. The test data is trained 2000 times so that the vector of each weight is located at the center of the input vector cluster, and the 6*6 competitive network is constructed. The network classifies or eliminates the screening of data, and obtains a healthy sample library that can be used to predict the running state of the machine in the future, achieving a good data cleaning effect.

Keywords: Self-organizing Map; Data cleaning

***Correspondence to Author:**
Guozhu Chen
College of Hydraulic & Environmental Engineering, China Three Gorges University, Yichang, 443002, China.

How to cite this article:
Song Yu, Guozhu Chen, Zhou Bao, Baogang Song. Abnormal data cleaning in thermal power plant based on self-organizing maps. Journal of Theoretical and Applied Sciences, 2019, 2:10

eSciPub LLC, Houston, TX USA.
Website: http://escipub.com/

1. Introduction

Thermal power plants have always played an important role in China's energy structure. In the thermal power production process, the tight coupling between the equipment, the complex system composition, and the high temperature, high pressure and high-speed rotation of the equipment, the thermal power plant equipment has been at a high failure rate. With the development of thermal power units to large capacity and high parameters, the impact of unit failures has increased significantly. Once the unit is shut down, it will not only cause large economic losses to the power plant itself, but also cause power grid accidents in serious cases, causing serious social consequences. Therefore, it is particularly urgent and important to strengthen the management of power generation equipment [1-2].

At present, power plants in power systems have established real-time monitoring information systems. These systems accumulate a large amount of power data. The thermal power plant needs to build a thermal power plant equipment intelligent diagnosis system based on the big data machine learning algorithm based on the real-time monitoring information of the power plant equipment and the historical data of the normal operation of the equipment. However, machine learning requires a large amount of reliability. Data, so cleaning the historical data to get a healthy sample database is a question worth exploring.

2. Problem analysis

How to filter and extract the device health sample database in a given massive data. For the big data processing problem, the most important thing is how to train the data and how to exclude the non-healthy data. First, establish the data sample taken and establish the parameter dimension. Because in real-time monitoring, the number of dimensions of the detected parameters is too large, and the time complexity for processing a single detection parameter is too high. At the same time, due to the correlation between the parameters, because the amount of data is large, the outliers belong to small samples, and should be eliminated. Abnormal data, the data to be inspected needs to be classified, and the simple clustering analysis error is too large. Therefore, it is necessary to establish a model to autonomously learn and statistically classify parameters, automatically find the inherent laws in the updated data and make the data autonomous. filter. Finally, the purpose of eliminating abnormal data is achieved. The specific flow chart of SOM data processing used in this paper is shown in Fig. 1:

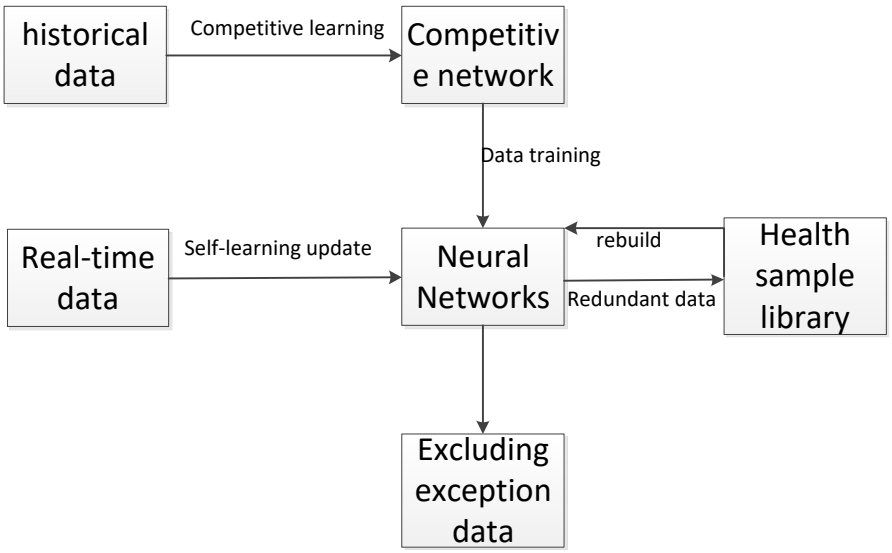


Fig. 1: Flowchart of data processing technology

3. Model establishment

Self-organizing feature mapping network (SOM), also known as Kohonen network, is composed of a neural network that is divided into different corresponding neuron regions when it accepts the external input mode. Each region has different response characteristics to the input mode, and the entire process is au-

tomatically completed. It does not require instructor guidance, and its characteristics are similar to those of the human brain. It can map any dimension input mode to one or two-dimensional discrete graphics at the output layer, and keep its topology unchanged[3].

SOM consists of a fully connected array of neurons, without the guidance of a tutor, high self-organization, and self-learning. It can obtain internal rules by repeatedly observing and comparing objective things, and classifying things with common characteristics. It can automatically find the intrinsic law and essential attributes in the sample, self-organizing adaptively change the network parameters and structure, determine the learning rules, and establish the training layout, so that each weight vector is located at the center of the input vector cluster. Once the SOM completes the training, it can be used to cluster the training data or other updated data. The specific process of the SOM training data is as follows:

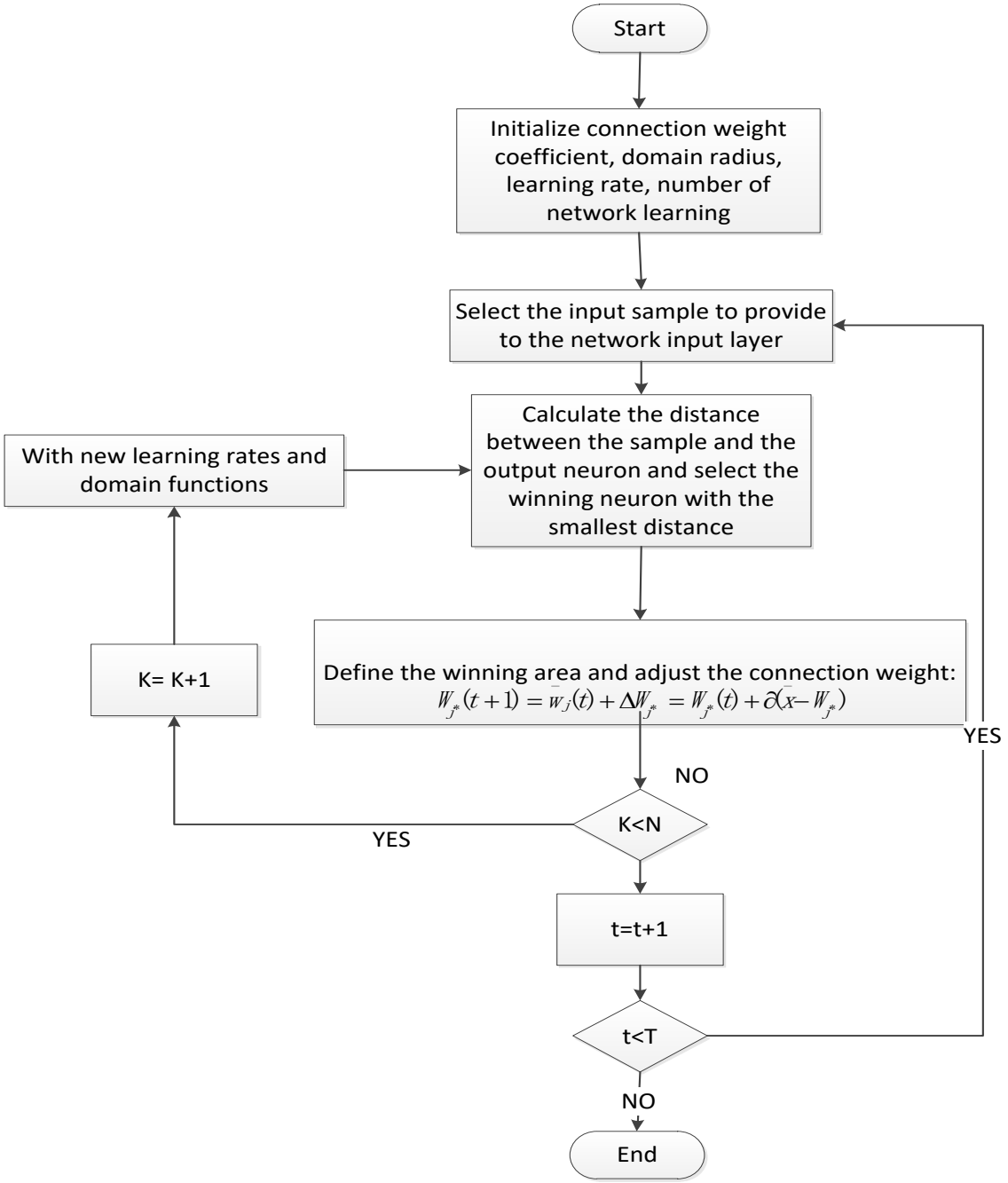


Fig. 2: SOM training data flow

The neuron construction pattern is shown in Fig. 3, which is as follows.

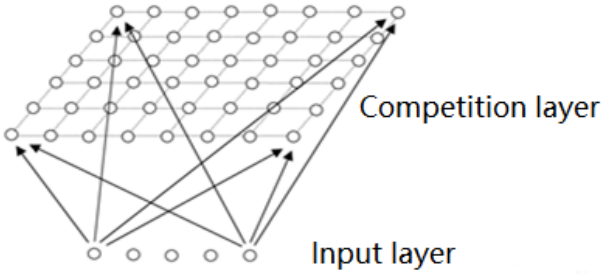


Fig. 3: Neuron construction pattern

Competitive learning is used for the training mode, and when the training sample is supplied to the data, the Euclidean distance between it and each weight is calculated. The neurons whose weight vector is most similar to the input become the best matching unit (BMU). The weight of the BUM in the varying SOM grid and the neurons adjacent to it will be adjusted toward the input vector. The amount from the BMU will decrease with time and distance (in the grid). The update formula for the neuron v with the weight Wv(s) is:

$$W_v(s+1) = W_v(s+1) + Q(u,v,s)\partial(s)(D(t) - W_v(s+1))$$

Where s is the step index, t is the training sample index, D(t) is the input vector, and u is the BMU index of D(t). is a monotonically decreasing learning coefficient, it can be repeatedly selected from (0, 1, 2, ..., T-1) of the system (T is the size of the training sample). It can also be randomly retrieved from the data set (Bootstrap sampling). All samples are trained here using systematic sampling.

The algorithm steps are as follows:

1. Vector normalization
   The current output mode vector X in the self-organizing neural network and the inner star weight vector $W_j(j=1,2,\mathrm{L}\ m)$ corresponding to each neuron in the competition layer are all normalized to obtain $\hat{x}$ and $\hat{W}_j$.

$$\hat{x} = \frac{x}{\|x\|}, \quad \hat{W}_j = \frac{W_j}{\|W_j\|}$$

2. Find winning neurons
   Compare the similarity between $\hat{x}$ and the inner star weight vector $W_j(j=1,2,\mathrm{L}\ m)$ corresponding to all neurons in the competition layer. The most similar neuron wins, the weight vector is $\hat{W}_j$
3. Adjustment of network output and rights
   According to the WTA learning rule, the winning neuron output is "1", otherwise it is 0, that is,

$$y_{i(t+1)} = \begin{cases} 1 & \mathrm{j}=\mathrm{j}^* \\ 0 & \mathrm{j} \neq \mathrm{j}^* \end{cases}$$

   Only the winning neuron has the right to adjust its weight vector $\hat{W}_j$, and its weight vector learning adjustment is as follows

$$\begin{cases} W_{j^*}(t+1) = \hat{W}_j(t) + \mathrm{V}W_{j^*} = \hat{W}_{j^*}(t) + \partial(\hat{x} - W_{j^*}) \\ W_j(t+1) = W_j(t) \end{cases}$$

   $\partial\ (0 < \partial \le 1)$ is the learning efficiency, and ∂ generally decreases as the learning progresses multidimensionally, that is, the degree of adjustment becomes smaller and smaller, tending toward the center of clustering.
4. Renormalization
   After the normalized weight vector is adjusted, the new vector obtained is no longer a unit vector. Therefore, the learned adjusted vector is renormalized

and looped until the learning rate decays to zero.

4. Model solution

In this paper, the 6441 data of the temperature difference between the upper and lower wall of the main part of the steam turbine (generator end), the temperature difference between the upper and lower walls of the medium pressure (the governor end) and the temperature difference between the upper and lower walls of the steam turbine are used as the sample data. Using the neural network library function of MATLAB, after modifying the learning times and establishing the generated dimensions, the training is performed by SOM neural network. The training results are shown in Fig. 4.

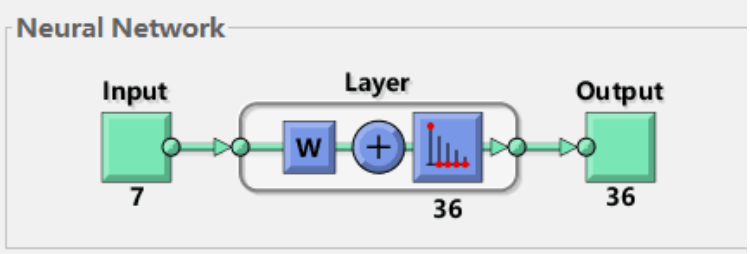


Fig. 4: Topology diagram of the neural network of the turbine body

The original data is numbered as 6441 groups. After trying to screen, a 6*6 competitive layer network is established. The neural network corporate topology map is as above, and the 6441 data is classified into 36 neurons (Fig. 5).

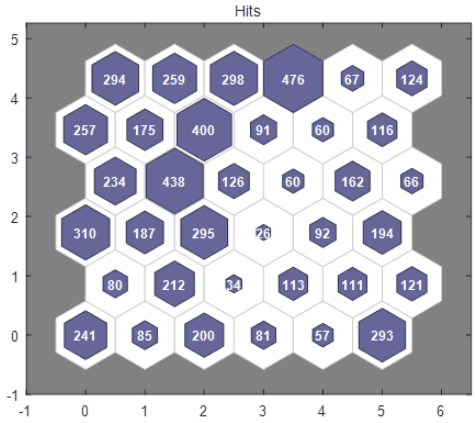


Figure 5: Neural network distribution of the turbine body part

Each hexagon in the figure represents a neuron, and the number in the hexagon represents the number of samples contained in the neuron. At the same time, the Euclidean distance between each neuron is shown in Fig. 6.

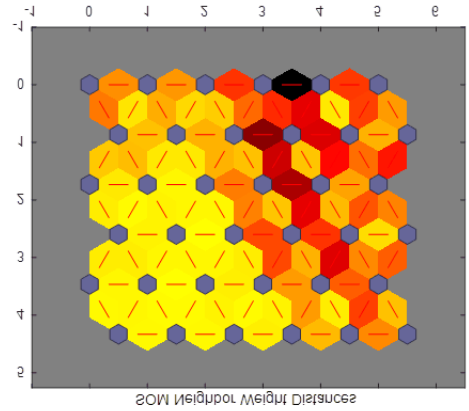


Fig. 6: Euclidean distance map of the neural network of the turbine body

It is easy to see from the above figure that the color of the neuron points at the 4th, 5th, 10th, 17th, 22nd, and 23rd is obviously deeper (the distance between the weights of the surrounding data fields is too large), and the focus is on the surrounding Euclidean distance. Excessive

neurons, the difference points are eliminated, and all the data in the distinct neurons are extracted.

5. Result analysis

Through the analysis of the collection of all the elements in the different neurons and the calculation of the distance between the corresponding neurons and their neighboring neurons, the original data is firstly removed, and the competitive network and neuron distribution obtained from the monitoring data of the turbine body are solved. It can be seen that the overall data distribution is more reasonable in the elements of the upper left neuron, while the weights in the middle and lower right corners are higher than the standard weight.

Since there is no detailed definition of the health indicators, we decided to remove the samples from the largest to the smallest to try to make the data as much as possible to remove more redundant values, first to eliminate the large range (Fig. 7).

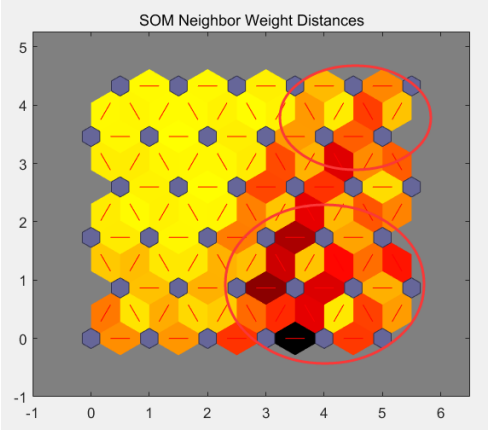


Fig. 7: End point removal of partial data results

Determine the knockout neurons as 1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 16, 17, 18, 22, 23, 24, 29, 30, 35, 36 neurons, total rejection There are 2,303 elements, and the remaining data of the remaining remaining health sample database is 4,097. In order to better observe the comparison between the fluctuation of the health sample database after the data is removed and the original data, it is placed on a line graph (Fig. 8).

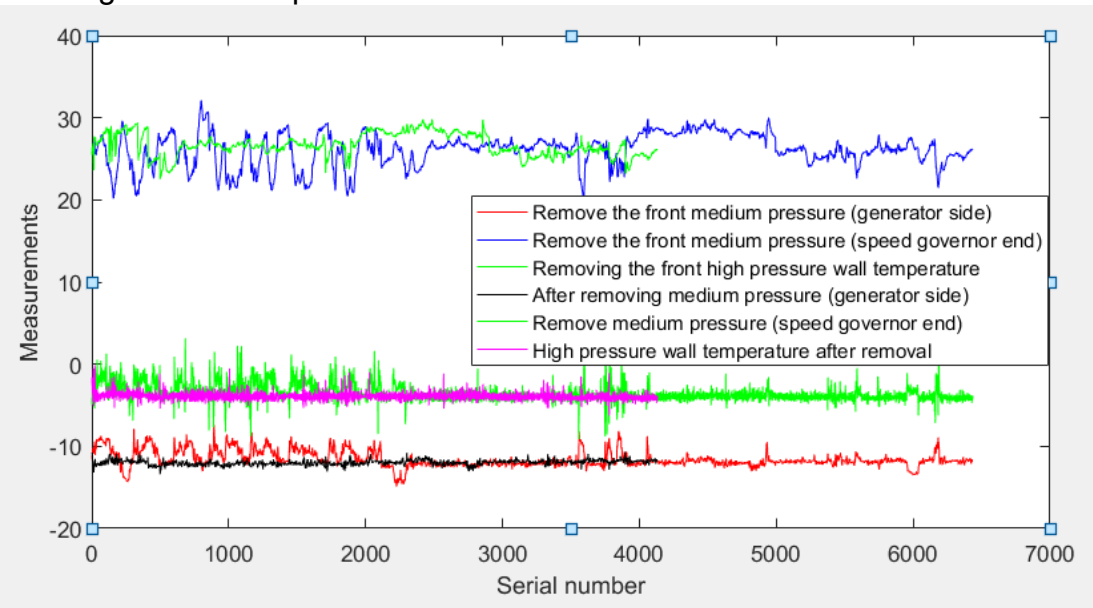


Fig. 8: Comparison of three indicators to remove outliers

It can be intuitively seen that although there is still local jitter, it is relatively normal in the actual situation that the data is relatively normal. It can be seen that the SOM model can play a good role in cleaning such data.

6. Conclusion

In this paper, the self-organizing feature map neural network (SOM) is used to clean the thermal power plant data, and the experiment is carried out by using MATLAB programming

with a large amount of data in the body part of the steam turbine. The satisfactory results are obtained. This is a healthy sample database for thermal power plants. Provides an idea to provide correct data protection for the automatic detection system of thermal power plants.

References

1. Minling Zhang, Zhaogan Chen, Zhihua Zhou. A Survey of SOM Algorithm, LVQ Algorithm and Its Variants[J]. Computer Science, 2002(07): 97-100.
2. Yufei Yue and Jianxu Luo. Application of an Improved SOM Neural Network in Fault Diagnosis of Wastewater Treatment[J]. Journal of East China University of Science and Technology, 2017,43(03):389-396.
3. Jinping Yu and Qin Xu. Data Cleaning Technology Based on SOM Network Clustering [J]. Science and Technology Plaza, 2005 (08): 59-61.