Improving K-means Clustering Method in Fault Diagnosis based on SOM Network

Anhua Chen
Hunan Provincial Key Laboratory of Health Maintenance for Mechanical Equipment, Hunan University of Science and Technology, Xiangtan, China
ahchen@hnust.edu.cn

Yang Pan and Lingli Jiang
Hunan Provincial Key Laboratory of Health Maintenance for Mechanical Equipment, Hunan University of Science and Technology, Xiangtan, China
604717678@qq.com, linlyjiang@163.com

Abstract—According to the problem of K value and initial cluster centers selection difficult on K-means clustering algorithm, form essential characteristics of the complex network, the fault samples can be abstracted into network nodes, and the connection between samples can be abstracted into edge, and then the network model of fault data can be established. Failure data network model is divided into several regions self-organizing feature map (SOM) network. K value can be determined from the maximum value which is selected in different division result by the use of community modularity at the same time. Complex network node correlation degrees can be calculated to select important nodes as initial clustering center, then by means of K-means clustering realizing clustering diagnosis. This study is applied to rolling bearing clustering the diagnosis examples and has good effect of fault diagnosis.

Index Terms—SOM network, Complex network, Community modularity, K-means clustering, Fault diagnosis

I. INTRODUCTION

Rolling bearing as a kind of general connection and transfer power parts in mechanical equipment plays an important role in almost any large equipment. When the equipment is running, Wear, fatigue, corrosion, overload and other reasons may cause the rolling bearing inner ring, outer ring and rolling elements damage. How to extract effective information in a number of fault information so that they could identify and distinguish state between normal and abnormal system (failure).because efficient troubleshooting is a problem which is very concerned by academia and industry. At present, the rotating machinery fault diagnosis research mostly take the method of neural network [1], Bayesian classification [2], support vector machine [3], the rough set theory [4], k-means clustering [5] and so on, all kinds of methods have different advantages and disadvantages. The final clustering results of K-means clustering algorithm, to some extent, depend on the choice of K value and initial cluster centers. While selecting different K value and initial cluster centers, clustering effect is different. Due to fault types are often unknown in the process of actual fault diagnosis and fault data is going to be divided into the number of categories, which user is not known. Under the condition of the number of clusters unknown to people, they often need to combine other algorithm get clustering number. That is, the value of K. Therefore, the certainty of K value is especially important in k-means algorithm. K-means clustering algorithm is heavily dependent on the initial clustering center choice. The traditional methods of selecting initial center such as: experience choose representative point, calculation method of center of gravity [6], "Density" method, before using the K a sample point as a representative point [7], etc. Experience choose representative point is very easy to create the clustering result into local optimal solution and even lead to wrong clustering result, computing center of gravity method choose clustering center is slow, etc.

In view of these disadvantages, in order to more rapidly and accurately find K value and initial clustering center of K-means clustering algorithm, this paper proposes a method that Improving K-means clustering method based on SOM network. Firstly, extracts feature vector, calculated node connection matrix, and fault data network model can be established. Network is divided into several data areas by using SOM network [8, 9]. According to complex network community structure modularity judges type number of faults for K-means algorithm looking for the K value, then the use of complex network degree of the sample data selection clustering center, finally, finishing clustering analysis by using K-means clustering algorithm. Through the example analysis shows that, Complex network community structure modularity is able to accurate select K values in k-means clustering algorithm. Compared with traditional selecting initial cluster center of K-means clustering algorithm method, In this paper, the use of complex networks degrees select cluster centers intuitive, clear and has a higher fault recognition rate, at the same time avoid the shortcomings of traditional methods.
Improving K-means clustering method based on SOM network verify a new method for the selection of K value and initial cluster center.

II. IMPROVED K-MEANS CLUSTERING ALGORITHM

K-means clustering algorithm is an indirect clustering method based on the similarity of sample between measures, belonging to the unsupervised learning method. K as a parameter in the process of this algorithm, n data samples are divided into k classes, in order to make within the cluster have high similarity and the degree of similarity between clusters is relatively low. According to the Euclidean distance calculation between each sample degree similarity, firstly, this algorithm chooses k objects, and each object represents a clustering center of mass. According to the distance between the object and each clustering center, the rest objects are assigned to they most similar clustering separately, repeat the process, until the n objects are assigned to complete.

K value is set by the user in k-means clustering algorithm, and the K value is difficult to directly identify in practical application, especially when the data amount large, how to identify the value of K will be a very great problem. When the K value select is different, the clustering result is different, through test K value to obtain in many methods, and these methods not have an accurate judgment basis. Therefore, in order to get the correct clustering result, it is important to determine the value of K. Initial cluster center select improper is very easy to cause the clustering result to fall into a local optimal solution or even lead to error in the result of clustering. In this paper, the initial clustering center is determined by the application of complex network node correlation degree [10], each node correlation degree is calculated in the network model, and chooses K important nodes as the initial clustering centers, at the same time network can clearly reflect relationship between the nodes, avoiding the defect of traditional methods.

A. Establish Fault Data Network Model

Complex network [11] is an abstraction and description about complicated system, any complex system contains a large number of units, when the units were abstracted into nodes, the relationship between units were abstracted for edge, and they can be research as a complex network. According to the complex network properties, the node can represent anything, we can know the nodes' relationship by the node analysis, different mechanical fault types' samples can be regarded as one complex network, and each sample is a node in the network. Through the analysis of the relationship between the complex networks' nodes can achieve the sample classification and diagnosis. Along with the intensive study of complex network as well as experiment, people found the nodes in the network, and the connection relationship between these nodes are often not be out of order, but contains some rules, most networks have a common nature -- community structure [11]. Community structure includes module class, crowd, group and other meanings, the nodes connection are more compact in internal community structure, but connection between the community structures are sparse. Community structure of complex network model [11] as shown in figure 1.

Through the field monitoring, we can collection dynamic fault information to extract the characteristic quantities of different fault types and component failure samples. Fault samples can be abstracted into network nodes, and the connection between samples can be abstracted into edge, so different fault types samples be seen as a network structure.

To collect the fault data composed of sample collection \(X = \{ x_1, x_2, \ldots, x_n \} \), each sample has \( p \) attribute, namely \( x_i = \{ x_{i1}, x_{i2}, \ldots, x_{ip} \}, (i=1,2,\ldots,n) \).

The relationship between \( x_i \) and \( x_j \) with similarity \( a_{ij} \in A \) representation, each data sample \( x_i \) as "node"", the link between the data samples can be expressed as "relations", then the data structure can be expressed into a weighted undirected network\[12\] \( G(X,A) \).

Comparison of similarity of different modes can be transformed into comparing distance of the two vectors. Generally speaking, \( a_{ij} \) is the function that the distance between sample \( x_i \) and \( x_j \). The principle of similarity function [12] design is that the network has good block structure (block inside as close as possible similarity, similarity degree between blocks is differences large). In this paper, it is defined as:

\[
    a_{ij} = \exp(-10 \times d_{ij}) 
\]

\[
    d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} 
\]

In the formula, \( d_{ij} \) using Euclidean distance metric, and the smaller \( d_{ij} \), the bigger \( a_{ij} \), this obviously shows that between \( x_i \) and \( x_j \) is similarity degree greater. Due to samples self-similarity has no meaning, this paper defined self-similarity to 0, that is to say, when \( i = j \), \( a_{ii} = 0 \). Due to the similarity between two nodes are equal, namely \( a_{ij} = a_{ji} \), so A is a symmetric matrix. The network connection matrix of n nodes:
\[
A = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
\]

So far, we can get fault data network model \( G(X, A) \).

B. Community Division based on SOM Network

Community modularity [13] is a measure of quality of the network partition metrics introduced by Newman. Fault data network is divided by SOM network C community. SOM network is that with self-learning function and it can simulate function of the human brain neural networks. SOM network is consisting of input layer and output layer. When SOM network accepts outside input, input data will be divided into different regions and each region has a different response characteristics of input mode, which different kinds of signal incentive is responded to different neurons that use the best way. This process is accomplished through unsupervised and adaptive. The purpose of clustering is that similar sample is classified as a class and not similar sample are separated, so we can realize pattern sample within the category of similar and inter-class separation. Because training sample of unsupervised learning do not contain expect output, so a sample input mode should belong to which category will not have any prior knowledge.

The SOM network structure [8] as shown in the Figure 2, which consists of input and output layers. Generally, the output layer is form of two-dimensional array, and input layer and output layer node are completed mutual connection.

\[\text{Figure 2. SOM network structure}\]

SOM network algorithm is as follows:

(1) Initialization

Each weight vector of SOM network output layer are endowed with smaller random number and normalized to obtain \( \hat{w}_j (j = 1, 2, \cdots , m) \). \( m \) is the number of output layer neuron. Then the initial outperformance neighborhood \( N'_j (0) \) and learning rate \( \eta \) initial values are established.

(2) Accept input

We are random take from an input mode and normalized in the training set, so can get \( \hat{X}^p (p = 1, 2, \cdots , n) \), \( n \) is the number of input layer neuron.

(3) Looking for winning node

We calculate dot product between \( \hat{X}^p \) and \( \hat{w}_j \) to find the largest dot product winning node \( j^* \). If the input mode without normalization, we should make use of formula (4) calculating Euclidean distance to find out the minimal distance between winning node.

\[d_j = \left\| \hat{X}^p - W_j \right\| = \sqrt{\sum_{j=1}^{m} (X_{j} - W_{j})^2}
\]

(4) Define winning neighborhood

We set \( j^* \) as weight adjustment domain in the t time. Generally, when initial neighborhood \( N'_j (0) \) is larger, \( N'_j (t) \) is shorter with training time in the training process.

(5) Adjust weights

Weights of all nodes will be adjusted in the winning neighborhood \( N'_j (t) \)

\[w_j (t+1) = w_j (t) + \alpha(t, N)(x_p^t - w_j (t)) \]

\( i = 1, 2, \cdots , n \) \( j \in N'_j (t) \)

In the formula, \( w_j (t) \) is a weight of neuron \( i \) in the \( j \) time. \( \alpha(t, N) \) is function that between the \( j \) neurons and winning neuron \( j^* \) topological distance N.

(6) End judgment

When learning rate \( \alpha(t) \leq \alpha_{\text{max}} \), we will finish training. Otherwise, it will turn to step (2) continue training. The above this self-organizing clustering complete in independent and under condition of without supervision.

According to SOM network classification results, we make \( x_p \) and \( x_q \) for get two subsets in the clustering obtain C set. The degree of similarity between two subsets is defined as \( e_{pq} \) [11]:

\[e_{pq} = \frac{\sum_{i \in X_p, j \in X_q} a_{ij}}{\sum_{i \in X_p, j \in X_q} a_{ij}^2} \quad p, q = 1, 2, \cdots , C
\]

When \( p = q \), \( e_{pq} \) is between subset within \( X_p \) a similarity measure. When \( p \neq q \), \( e_{pq} \) is subset between \( X_p \) and \( X_q \) similarity measure. We can define a symmetric matrix \( E = (e_{ij}) \) of \( k \times k \) dimension. At the same time, we draw on the concept of complex network community modularity that it is defined as \( Q \) :

\[Q = \sum_{p=1}^{C} e_{pp} - \sum_{p=1}^{C} \left( \sum_{q=1}^{C} e_{pq} \right)^2
\]

In this formula, the \( \sum_{p=1}^{C} e_{pq} \) reflect community internal node connection and \( \sum_{p=1}^{C} \left( \sum_{q=1}^{C} e_{pq} \right)^2 \) reflect between
community and community node connection. Obviously, the bigger \( \sum_{p=1}^{c} e_{pq} \), the smaller \( \sum_{p=1}^{c} (\sum_{q=1}^{c} e_{pq})^2 \), \( Q \) value larger shows community structure more obvious.

We can with fault data as complex network structure, so that each community can be seen as each type of fault. It is used of Modularity \( Q \) value bigger show the community structure is more obvious, so as to determine community number, which the result can for k-means clustering algorithm provide for K value and solve the K value acquisition difficulties problem.

C. Degree of Complex Network

The degree of complex network [14, 15] is a concept about node. The correlation degree between two adjacent nodes is determined by their share of edge, the node degree of \( i \) with \( k_i \) said, which defined as the number of branches of the side connected with node \( i \). Visual point of view, the greater the degree of a node is, the more important it is in the k-means algorithm.

We use node degree selection method to select the cluster center in the k-means algorithm. This paper extract characteristic parameter for the normal, inner ring and outer fault three fault types and component fault samples set. Each sample can be abstracted into network nodes, and the connection between samples can be abstracted into edge. According to the definition of the judgment factor, we will be greater than or equal to the judgment factor as 1, showing that two nodes have connection, less than it to 0, showing that there has been no connection. Complex network structure can be abstracted into a diagram. In this diagram, we can calculate each node degree, and then find out the biggest k node degree. This is what we are looking for clustering center.

Thus, Improving K-means clustering diagnosis method based on SOM network described as follows:

Step1: Collection of fault data, extraction feature component failure sample set;
Step2: Input fault data to construct the similarity matrix \( A \), fault data network model \( G(X,A) \) is established;
Step3: Fault data network is divided into C community makes use of SOM network;
Step4: Calculation of network model community modularity for k-means clustering algorithm provides the K value;
Step5: Select appropriate judgment factor to calculate the degree of each node in the network model;
Step6: Find out K the maximum node degree as the initial cluster centers of K-means clustering algorithm;
Step7: K-means clustering to realize fault diagnosis.

III. ROLLING BEARING FAULT PATTERN RECOGNITION EXAMPLE

Take rolling bearing fault diagnosis for example, Spectra Quest Company (USA) machinery fault simulator table as rolling bearing experiment equipment, in order to verify the validity of this method above. As shown in figure 4, the rotor-bearing system is driven 3 HP frequency conversion motor. The right side with rolling bearings fault, the left side is the normal bearing in the same condition. We set rotation frequency of the shaft for 30Hz and sampling frequency for 12 KHz in the process of experiment, and using the Austrian Dewetron company DEWE-16 channel high-precision test system. The mechanical fault simulator experiment as shown in figure 4.

![SOM network](image-url)

Figure 3. The flow chart of improving K-means clustering diagnosis method based on SOM network
We collect normal bearing, inner ring fault and outer ring failure vibration signal of three states in the same condition. Collected vibration signal are processed with empirical mode decomposition (EMD) \([16, 17]\). The inner ring fault signal by EMD as shown in the figure.

According to formula (8), (9) and (10), we can calculate the energy entropy value that EMD get intrinsic mode function of time sequence \(i = 1, 2, \cdots n\).

\[
E = \sum_{i=1}^{n} x_i^2
\]

\[
H_{EN} = -\sum_{i=1}^{8} p_{emd_i} \log p_{emd_i}
\]

\[
p_{emd_i} = \frac{E_{emd_i}}{\sum_{i=1}^{8} E_{emd_i}}
\]

In this paper, we extract 50 samples from each fault type to form 150 x 1 sample set. Normal bearings, inner ring fault and outer ring fault samples be abstracted into a complex network, each sample as complex network a node, connection between sample and sample is abstract for edge, so complex network model \(G(X, A)\) with 150 nodes. First, according to formula (1) and (2) we establish a connection matrix \(A\) that it is a symmetric matrix of 150x150, and all zeros are on the main diagonal, As is shown in table 1.

Failure data network model is divided into several regions by SOM network. Its division results as shown in figure 6.

According to formula (6) and (7), when \(k\) take different value, we can calculate the corresponding modularity value of \(Q\). The results are shown in Figure 7.

The figure 7 shows that when divide three communities, modularity \(Q\) value is the maximum, and we think that the ideal community division should be clustered into three categories. So we can determine contain three kinds of fault types in the failure data network model. This result is consistent with the experiment. It proves that this paper proposed method can find \(K\) value for K-means clustering algorithm.

According to the \(K\) value offered above, randomly select \(K\) initial cluster centers to K-means clustering algorithm. Take 0-50 as normal bearing, 51-100 as inner ring fault and 101-150 as outer ring fault. The results are shown in figure 8.

It is known that the clustering results are poor in the figure 8. In the paper, we select initial cluster center by the complex networks degree. To calculate the size of the degree of each node in the network, setting a factor \(\phi\), when the connection matrix \(A\) value is larger than \(\phi\), it is set to 1 and smaller than \(\phi\) is set to 0. Because factor \(\phi\) has no definite standard, therefore, we begin to take \(\phi = 0.5\) steps of 0.05 increasing. Through 10 times test, we conclude that when \(\phi = 0.75\), the size of each node degrees can be clearly distinguished. Greater than \(\phi = 0.75\) is set to 1, less than \(\phi = 0.75\) is set to 0. The size of degree of each sample is shown in figure 9.

<table>
<thead>
<tr>
<th>(a)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.56159</td>
<td>0.85255</td>
<td>0.13522</td>
<td>0.21491</td>
<td>0.9982</td>
<td>0.92687</td>
<td>0.99999</td>
</tr>
<tr>
<td>2</td>
<td>0.56159</td>
<td>0</td>
<td>0.87833</td>
<td>0.65122</td>
<td>0.79393</td>
<td>0.47669</td>
<td>0.79114</td>
<td>0.56425</td>
</tr>
<tr>
<td>3</td>
<td>0.85255</td>
<td>0.87833</td>
<td>0</td>
<td>0.35685</td>
<td>0.49334</td>
<td>0.77837</td>
<td>0.98478</td>
<td>0.85467</td>
</tr>
<tr>
<td>4</td>
<td>0.13522</td>
<td>0.65122</td>
<td>0.35685</td>
<td>0</td>
<td>0.97</td>
<td>0.10054</td>
<td>0.2733</td>
<td>0.13642</td>
</tr>
<tr>
<td>5</td>
<td>0.21491</td>
<td>0.79393</td>
<td>0.49334</td>
<td>0.97</td>
<td>0</td>
<td>0.16553</td>
<td>0.39452</td>
<td>0.21658</td>
</tr>
<tr>
<td>6</td>
<td>0.9982</td>
<td>0.47669</td>
<td>0.77837</td>
<td>0.10054</td>
<td>0.16553</td>
<td>0</td>
<td>0.8677</td>
<td>0.98919</td>
</tr>
<tr>
<td>7</td>
<td>0.92687</td>
<td>0.79114</td>
<td>0.98478</td>
<td>0.2733</td>
<td>0.39452</td>
<td>0.8677</td>
<td>0</td>
<td>0.92846</td>
</tr>
<tr>
<td>8</td>
<td>0.99999</td>
<td>0.56425</td>
<td>0.85467</td>
<td>0.13642</td>
<td>0.21658</td>
<td>0.99919</td>
<td>0.92846</td>
<td>0</td>
</tr>
</tbody>
</table>

© 2013 ACADEMY PUBLISHER
As shown in Figure 9, we take the largest number of nodes degrees respective for $x_{29}$, $x_{31}$, $x_{125}$, namely as
the initial clustering center. Then K-means clustering, the cluster effect is shown in Figure 10.

Basically, all kinds of fault can be distinguished. This paper proposes that complex network degree can select the initial cluster centers for K-means clustering algorithm, and can be applied in rolling bearing pattern recognition. In this paper, we compute accuracy rate of improve K-means clustering method and K-means clustering methods, as shown in Table II: Table II shows that improving K-means clustering diagnosis method based on SOM network in normal, inner ring failure, outer ring failure, the overall accuracy rate are higher than the traditional K-means clustering algorithm. This study proves that the algorithm proposed in the paper has good fault diagnosis effect and it can be applied in the fault diagnosis.

V. CONCLUSION

According to the disadvantages of k-means clustering algorithm in selecting K value and initial clustering center, the paper expects to take advantage of the characteristics of complex network to abstract fault samples into the network nodes and the connection between samples is abstracted into edge, then the network model of fault data can be established. Failure data network model are divided into several regions by using SOM network, and then we can make use of complex network modularity to determine the corresponding classification results in different categories. When Q value the maximum, according to the classification result for K-means clustering algorithm determine K value. Complex network degree is used to calculate the size of the degree of each node, and select K clustering initial center. Compared with the traditional K-means clustering method applied to rolling bearing fault diagnosis examples, it proves that our method has higher accuracy rate. The improves K-means clustering method based on SOM network provides a new way for selecting K value and initial clustering center.

ACKNOWLEDGEMENTS

This work is supported by the national natural science foundation of China (51175169, 51105138), the national high technology research and development program items (2012AA041805), the Pre-research project (813040302), Hunan university of science and technology innovation fund project (S122015) and the aid program for science and technology innovative research team in higher educational institutions of Hunan province.

REFERENCES


Anhua Chen, male, born in May, 1963, in Heng-yang city of Hunan province, received PHD in 1997 from Central South University, research fields are health maintenance and control mechanical fault dynamics.

He is currently a professor and vice-principal in the Hunan University of Science and Technology, China, mainly engaged in scientific research, teaching and student management. His other works include “dynamics theory and method of vibration diagnosis” (Beijing, Mechanical Industry Press, December 2002), “Numerical investigations on dynamic transmission error and stability of a geared rotor-bearing system”, (Chinese, Journal of Mechanical Engineering, 2004), and so on. His research interests include mechanical dynamics, fault diagnose, dynamic test and signal process and equipment maintenance technology.

Prof. Chen is the Senior Member of the Chinese Mechanical Engineering Society and director of fault diagnosis branch of Chinese Vibration Engineering in recent years. He is presiding more than 5 research projects and had published more than 80 academic papers.

Yang Pan, male, born in 1986. He is a graduate student in Key Laboratory of Health Maintenance Equipment of Hunan Province, Hunan University of Science and Technology, Xiangtan, China. His main research direction is mechanical fault diagnosis and equipment health maintenance technology.

Lingli Jiang, female, born in 1981. She received MSc from Hunan University of Science and Technology in 2007. She obtained her PHD degree from Central South University in 2010. She has been working at Hunan University of Science and Technology since 2010. Her research interest is mechanical equipment fault diagnosis and signal process. Her works include: “Degradation assessment and fault diagnosis for roller bearing based on AR model and fuzzy cluster analysis”. (Shock and Vibration, 2010), and so on.