2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science

# Research on Application of AP Clustering Algorithm in Fault Diagnosis

Juncai Lin
Computer Science and Technology
Wuhan University of Technology
Wuhan, China
e-mail: 1978201677@qq.com

Qing Yang
Computer Science and Technology
Wuhan University of Technology
Wuhan, China
e-mail: 2816189203@qq.com

***Abstract*—K-means and k-centers clustering algorithms need to pre-configure the number of clusters and its clustering granularity are rough. To solve this problem, the AP clustering algorithm was applied to mechanical fault diagnosis field. The EEMD and approximate entropy theory were used to extract fault features from fault data set. Then the AP algorithm was used to discover the fault pattern from extracted fault features. Finally the new sample's fault type was diagnosed according to the clustering result. Experimental results showed that AP clustering algorithm could effectively improve the accuracy of clustering and improve the accuracy of fault diagnosis without pre-configuring the number of cluster centers.**

***Keywords- data mining; AP clustering; fault diagnosis***

## I. INTRODUCTION

Clustering methods in data mining do not rely on existing prior knowledge and can effectively classify unlabeled data. All objects in a cluster are similar to each other, and the objects between clusters and clusters are different from each other. Therefore, the clustering methods are widely used in the field of mechanical fault diagnosis, especially in the field of fault diagnosis of rotating machinery such as bearings, rotors and engine systems. Clustering algorithms such as FCM, k-means, k-centers, GK, and GG are commonly used in fault diagnosis [1-5]. The FCM clustering algorithm is an improvement of the ordinary C-means algorithm. The FCM-based fault diagnosis method has a strong generalization ability and is suitable for data with a spherical distribution. The k-means algorithm is simple and convenient for its calculation. The advantages of compact clusters and obvious separation between clusters and clusters are widely used in various fields, but they are easily affected by noise points. The k-centers algorithm solves the problem that the k-means algorithm is sensitive to noise points. GK clustering algorithm obtains the objective function based on covariance matrix, and can find irregular clusters. So it is suitable for analyzing irregularly distributed data sets. The GG clustering algorithm measures the similarity between data samples by introducing the maximum likelihood distance, which is suitable for data with any shape distribution.

However, the above clustering algorithms all need to pre-configure the number of clusters. For example, the popular k-centers clustering algorithm starts from the initial set of randomly selected samples and iteratively refines the set to reduce the sum of the squared errors. So k-centers clustering algorithm is very sensitive to the initial selection of the sample. Usually, it needs to randomly select different initial sample sets several times, and then iteratively executes the algorithm to find a good solution from multiple clustering results. However, the clustering algorithm can achieve sufficiently good results only when the number of clusters is small and at least one random initial sample set is close to the optimal result. Therefore, these algorithms cannot guarantee convergence to the optimal clustering results. In the actual fault diagnosis application, the data set composed of fault features is not always a spherically distributed data set. If the number of fault categories is directly used as the cluster center number for clustering, the clustering granularity will be rough so that the accuracy of fault diagnosis will be reduced significantly.

The AP clustering algorithm [6] uses all the data objects in the data set as candidates for the cluster center in the clustering process, and uses the similarity between the data objects for cluster analysis, and does not require the number of cluster centers to be set in advance. The clustering result of AP clustering algorithm is fine-grained, and the number of cluster clusters is independent of the number of sample categories. It is only related to the preference parameter in the algorithm. When the value of the preference parameter is large, the number of cluster clusters is large. It will be improved along with the finer granularity of clustering. Therefore, this paper introduced the AP algorithm into the field of mechanical fault diagnosis. The AP clustering algorithm does not need to pre-configure the number of cluster centers and the fine granularity of the clustering results can improve the accuracy of fault pattern recognition.

## II. AP CLUSTERING ALGORITHM

AP clustering algorithm is a clustering algorithm based on "information transfer" between data objects. Unlike k-means algorithm or k-NN algorithm, AP algorithm does not need to determine the number of clusters before running the algorithm. It treats all data points as potential exemplar.

The AP clustering algorithm treats each data point as a node in the network and recursively passes real-value messages along the edges of the network until a good set of examples (cluster centers) and corresponding clusters appear.

For each data point in the data sample set $X=(x_1,x_2,\cdots,x_n)$, a matrix $S=(s(i,j))$ is used to represent the similarity between the data points. $s(i,j)$ indicates the similarity between the data point $x_i$ and the data point $x_j$. The larger the $s(i,j)$ value, the closer the distance

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00061

between the data point $x_i$ and the data point $x_j$ . $s(i,j) > s(i,k)$ indicates the distance between the data point $x_i$ and the data point $x_j$ is closer than the distance between the data point $x_i$ and the data point $x_k$ . $s(i,i)$ represents the probability of the data point $x_i$ becoming the center of the cluster. The value is larger, indicating that the data point $x_i$ is more likely to be selected as the center of the cluster. When the algorithm is initialized, $s(i,i)$ is the value given by the user, that is, the preference parameter. The size of the value will affect the number of clusters at the end of the algorithm. The larger the value, the greater the number of cluster clusters finally obtained. The similarity between data points is calculated as in

$$s(i,j) = -\left\|x_i - x_j\right\|^2, i \neq j \tag{1}$$

The AP clustering algorithm is actually a process of iteratively updating the matrix $R = (r(i,j))$ and the matrix $A = (a(i,j))$ . In the matrix $R$ , $r(i,k)$ represents the possibility of data point $x_k$ as data point $x_i$ 's cluster center. In the matrix $A$ , $a(i,k)$ represents the possibility of the data point $x_i$ choosing data point $x_k$ as its cluster center. The calculation equation of the matrix $R$ is shown in (2), and the calculation equation of the matrix $A$ is shown in (3).

$$r(i,k) = s(i,k) - \max_{k' \neq k} \left\{ a(i,k') + s(i,k') \right\} \tag{2}$$

$$\begin{cases} A(i,k) = \min\left\{0, R(k,k) + \sum\limits_{i' \notin \{i,k\}} \max\left\{0, R(i',k)\right\}\right\}, i \neq k \\ A(i,k) = \sum\limits_{i' \neq k} \max\left\{0, R((i',k))\right\}, i = k \end{cases} \tag{3}$$

In order to avoid oscillations and speed up the convergence of the iterative results in the message iteration process, the AP algorithm introduces a damping factor $\lambda$ , where each piece of information is set to a multiple of the value of its last iteration update and times the updated value of this information. The damping factor is often set to a real number between 0 and 1[6]. The equations for updating the matrix $R$ and $A$ are shown in equation (4) and equation (5).

$$R_t = (1-\lambda) R_t + \lambda R_{t-1} \tag{4}$$

$$R_t = (1-\lambda) R_t + \lambda R_{t-1} \tag{5}$$

The iterative process terminates when the number of iterations reaches the maximum number of iterations or the iterative process converges.

According to the equation (6), the clustering result is obtained. When $i = j$ , the data point $x_i$ is the center of the cluster. The process of AP algorithm is shown in Algorithm 1

$$\arg\max_{1 \le j \le N} \left[ R(i,j) + A(i,j) \right] \tag{6}$$

**Algorithm 1 AP**

**Inputs:** Similarity matrix $S$ , convergence iteration times $t_{con}$ , maximum iteration times $t_{\max}$ , damping factor $\lambda$ , preference parameter $p$

**Outputs:** cluster center $A$ , clusters $C$

**Steps:**

1.initialize the similarity matrix $S$ according to the preference parameter and equation (1);

2.initialize the matrix $R$ and the matrix $A$ according to $S$ , equation (2) and equation (3);

3.update the matrix $R$ and the matrix $A$ according to equation (4) and equation (5);

4.go to step 3 until reaching the maximum iteration times $t_{\max}$ or the iterative process converges;

5.get cluster centers and clusters according to equation (6).

## III. PROCESS OF FAULT DIAGNOSIS

Fault feature extraction and fault pattern recognition are the keys to the fault diagnosis. Fault feature extraction is the process of extracting features that can characterize the machine state from monitoring signals. Fault pattern recognition is the process of classifying different fault samples according to the extracted fault features. The fault diagnosis model based on AP clustering algorithm is shown in Fig. 1.

The fault diagnosis model based on AP clustering includes fault feature extraction module, cluster analysis module and fault diagnosis module. Firstly, the fault features are extracted from the fault samples and are composed to the fault feature vectors. Then, the clustering algorithm is used to obtain the fault pattern from the fault feature vectors. Finally, fault diagnosis is performed on the new fault samples according to the feature vectors and clustering results.

EEMD and approximate entropy theory have been widely used to extract fault feature in the fault diagnosis[7-9]. This paper also uses EEMD and approximate entropy theory. The detail information can be seen in [7-9].

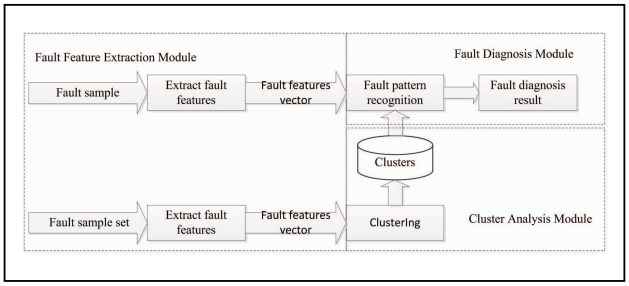


Figure 1. fault diagnosis model based on AP clustering algorithm.

The fault diagnosis model is used to identify the fault samples. This is the purpose of the entire fault diagnosis process. The fault diagnosis process is shown in Algorithm 2.

**Algorithm 2 Fault Diagnosis Process**

**Inputs:** Clusters $C$ , fault feature vectors $V$ , sample $x=(x_1,x_2,\cdots,x_n)$

**Outputs:** sample's category $c$

**Steps:**

1. extract sample's feature vector by EEMD and approximate entropy theory;
2. choose the closest cluster by comparing the distances between the sample and cluster centers;
3. compare the distance of the sample and the closest cluster center and the max distance inside the closest cluster. If the distance of the sample and the closest cluster center is bigger, then set "unknown" as sample's category. Otherwise, set the most category of the cluster as sample's category

In step 3, the maximum distance in the cluster is used as a threshold, and the samples that do not belong to any cluster are classified as "unknown categories".

## IV. Experiment Analysis

### A. Evaluation Criteria

The evaluation criteria used in this paper include three quality standards, namely Sum of Similarities (SS), clustering accuracy (Accu), and Normalized Mutual Information (NMI).

The goal of a distance-based clustering algorithm is to find cluster centers so that the sum of the distances between each data point and the center of the cluster is the smallest. Therefore, SS is an important criterion. The greater the SS, the better the clustering effect. The equation for calculating SS is shown in equation (7).

$$SS=\sum_{i=1}^{N}s(i,c_i) \tag{7}$$

Where $s(i,c_i)$ represents the similarity between the data point $x_i$ and the center to which it belongs.

NMI is a criterion of information theory. It is commonly used in clustering to measure the similarity of clustering results. The equation for the calculation of NMI is shown in equation (8).

$$NMI=\frac{I(c,c^*)}{\sqrt{H(c)H(c^*)}} \tag{8}$$

$I(c,c^*)$ represents the mutual information between the clustering result and the actual cluster label. $H(c)$ represents the information entropy of the clustering result. $H(c^*)$ represents the information entropy of the actual cluster label.

The clustering accuracy intuitively reflects the quality of the clustering results. The calculation equation is shown in equation (9).

$$Accu=\frac{\sum_{i=1}^{N}\delta(c_i,c_i^*)}{N} \tag{9}$$

Where $c_i$ represents the actual cluster label and $c_i^*$ represents the clustering result label.

### B. Experimental Data Set

The bearing data of Case Western Reserve University was used as the experimental data sets[10]. This data set includes fault data for both the drive and fan end bearings. This paper used the drive end rolling bearing fault data for experiments. The bearings are double-sided sealed bearings. Bearings were treated with a single point of EDM damage and the sampling frequency was 12 kHz. Sampling data is divided into four categories, which are Normal (NR), Rolling Fault (BF), Inner Race Fault (IRF), and Outer Race Fault (ORF).

In this paper, two sets of data with 0.1778mm damage diameter, 1797RPM bearing operating speed, 0.1778mm damage diameter, and 1772RPM bearing operating speed were used as experimental data sets.

The data were randomly divided into two parts in a ratio of 2:1. The one part formed a cluster analysis sample set.The other part formed a fault diagnosis sample set. The cluster analysis sample set was used to verify the clustering accuracy of the AP clustering algorithm in the cluster analysis module. The fault diagnosis sample set was used to verify the diagnostic accuracy in the fault diagnosis module.

In the dataset with 0.1778mm damage diameter and 1797RPM bearing working speed, the total number of experimental data samples was 296, including 119 NR status samples and 59 BF, IRF, and ORF status samples. The number of sample points in each sample was 2048. The experimental data details are shown in Table I.

TABLE I. Fault sample data set with a working speed of 1797 RPM

| Status | Damage Diameter (mm) | Working Speed (RPM) | Sample Size | Cluster Analysis Sample Size | Diagnosis Sample Size |
|---|---|---|---|---|---|
| NR | 0 | 1797 | 119 | 80 | 39 |
| BF | 0.1778 | 1797 | 59 | 40 | 19 |
| IRF | 0.1778 | 1797 | 59 | 40 | 19 |
| ORF | 0.1778 | 1797 | 59 | 40 | 19 |

TABLE II. Fault sample data set with a working speed of 1772 RPM

| Status | Damage Diameter (mm) | Working Speed (RPM) | Sample Size | Cluster Analysis Sample Size | Diagnosis Sample Size |
|---|---|---|---|---|---|
| NR | 0 | 1772 | 236 | 157 | 79 |
| BF | 0.1778 | 1772 | 59 | 40 | 19 |
| IRF | 0.1778 | 1772 | 59 | 40 | 19 |
| ORF | 0.1778 | 1772 | 59 | 40 | 19 |

In the data set with a damage diameter of 0.1778 mm and a bearing working speed of 1772 RPM, the total number of experimental data samples is 413, including 236

are NR status samples, and 59 are BF, IRF, and ORF status samples. The number of sample points in each sample is 2048. The experimental data details are shown in Table II.

### C. Experimental Results

Since the clustering result of the k-means algorithm depends on the selection of the initial clustering center. Each time the algorithm runs, the results are different. Therefore, the k-means algorithm was run 10 times. And the average values were taken as the final clustering result. In order to verify the clustering granularity of the AP clustering algorithm, the AP clustering algorithm was run first. And then the category number of fault samples and the number of cluster categories of the AP clustering results were set as the initial number of the cluster center in the k-means algorithm.

The clustering results when the bearing working speed is 1797 RPM are shown in Table III. And the clustering results when the bearing working speed is 1772 RPM are shown in Table IV. In the table, "n" represents the number of the cluster centers.

In Table III and Table IV, it can be seen that the k-means algorithm performed better when clustering center data is 9 then 4. However, the AP clustering results performed best in SS, NMI, and the average clustering accuracy. That showed AP clustering results could get higher accuracy than k-means algorithm in fault pattern recognition.

The diagnostic results when the bearing working speed is 1797 RPM are shown in Table V. And the diagnostic results when the bearing working speed is 1772 RPM are shown in Table VI. In the table, "n" represents the number of the cluster centers.

TABLE III. CLUSTERING RESULTS WITH A WORKING SPEED OF 1797 RPM

| algorithm | n | Accuracy(%) | | | | Avg | SS | NMI |
|---|---|---|---|---|---|---|---|---|
| | | *NR* | *BF* | *IRF* | *ORF* | | | |
| AP | 9 | 97.5 | 100 | 100 | 97.5 | 98.5 | 5.161 | 0.986 |
| k-means | 4 | 97.5 | 90 | 87.5 | 92.5 | 93 | 14.22 | 0.891 |
| k-means | 9 | 98.75 | 87.5 | 90 | 87.5 | 94.5 | 6.526 | 0.936 |

TABLE IV. CLUSTERING RESULTS WITH A WORKING SPEED OF 1772 RPM

| algorithm | n | Accuracy(%) | | | | Avg | SS | NMI |
|---|---|---|---|---|---|---|---|---|
| | | *NR* | *BF* | *IRF* | *ORF* | | | |
| AP | 15 | 96.81 | 95 | 100 | 97.5 | 97.11 | 5.70 | 0.965 |
| k-means | 4 | 94.9 | 95 | 90 | 87.5 | 93.14 | 18.3 | 0.896 |
| k-means | 15 | 96.17 | 95 | 92.5 | 92.5 | 94.94 | 6.82 | 0.942 |

TABLE V. DIAGNOSTIC RESULTWITH A WORKING SPEED OF 1797 RPM

| algorithm | n | correct number | Accuracy(%) | | | | Avg |
|---|---|---|---|---|---|---|---|
| | | | *NR* | *BF* | *IRF* | *ORF* | |
| AP | 9 | 92 | 94.87 | 94.73 | 94.73 | 100 | 95.83 |
| k-means | 4 | 84 | 89.74 | 84.21 | 78.94 | 94.73 | 87.5 |
| k-means | 9 | 87 | 89.74 | 94.73 | 84.21 | 94.73 | 90.63 |

TABLE VI. DIAGNOSTIC RESULT WITH A WORKING SPEED OF 1772 RPM

| algorithm | n | correct number | Accuracy(%) | | | | Avg |
|---|---|---|---|---|---|---|---|
| | | | *NR* | *BF* | *IRF* | *ORF* | |
| AP | 15 | 131 | 97.46 | 89.47 | 100 | 94.73 | 96.32 |
| k-means | 4 | 128 | 94.93 | 89.47 | 94.73 | 94.73 | 94.11 |
| k-means | 15 | 130 | 96.2 | 94.73 | 89.47 | 99.12 | 95.59 |

In Table V and Table VI, it also can be seen that the k-means algorithm performed better when clustering center data is 9 then 4 in the fault diagnosis. However, the AP clustering results performed best in the average diagnostic accuracy. That showed AP clustering results could get higher accuracy than k-means algorithm in the fault diagnosis.

Compared with the bearing working speed of 1797 RPM, the AP clustering accuracy rate and diagnostic accuracy rate when the bearing working speed is 1772 RPM were higher. This was because there were more fault samples in the data set when the bearing working speed is 1772 RPM, AP clustering algorithm could excavate more information in the fault samples.

## V. CONCLUSION

In this paper, AP clustering algorithm is applied to mechanical fault diagnosis. Firstly, EEMD and approximate entropy theory were used to extract fault feature from the fault samples and compose the feature vectors. Then the feature vectors were used as the input of the AP clustering algorithm to get the fault pattern. Finally, fault diagnosis was performed according to the clustering results. Experimental results showed that AP clustering algorithm could effectively improve the accuracy of the fault pattern recognition without the need to pre-configure the number of cluster centers. So that, the accuracy of fault diagnosis was improved.

## REFERENCES


[1] Zhang Shuqing, Sun Guoxiu, Li Xinxin, and Jian Xiong, "Study on mechanical fault diagnosis method based on LMD approximate entropy and fuzzy C-means clustering", Chinese Journal of Scientific Instrument, Vol 34, 2013, pp. 714-720.

[2] Liu Jiangei, "Application of clustering algorithm in rotor fault diagnosis", Xi'an Technological University, 2015.

[3] Zhou Yunlong, Wang Suobing, and Zhao peng, "Analysis of fan vibration based on improved k-means clustering algorithm". Journal of Vibration Measurement & Diagnosis, Vol 32, 2012, pp. 437-440.

[4] .Wang Shutao, Li Liang, Zhang Shuqing, and Sun Guoxiu, "Mechanical fault diagnosis method based on EEMD sample entropy and GK fuzzy clustering", China Mechanical Engineering, Vol 24, 2013, pp. 3036-3040.

[5] Zhang Shuqing, Bao Hongyan, Li Pan, Li Xinxin, and Jiang Wanlu, "Fault diagnosis of rolling bearings based on RQA and GG clustering", China Mechanical Engineering, Vol 26, 2015, pp. 1385-1390.

[6] Frey B J, Dueck D, "Clustering by Passing Messages Between Data Points", Science(New York), Vol 315,Jan. 2007, pp. 972-976, doi: 10.1126/science.1136800.

[7] Qi Peng, Fan Yugang, and Feng Zao, "Study on fault diagnosis method based on SSVD and EEMD", Transducer and Microsystem Technologies, 2018.

[8] Zhou zhi, Zhu Yongsheng, Zhang Youyun, Zhu Chuanfeng, and Wang Peng. "Adaptive fault diagnosis of rolling bearing based on EEMD and demodulated resonance". Journal of Vibration and Shock, Vol 32, 2013, pp. 76-80.

[9] Huang Youpeng, Zhao shan, and Xu Fan, "Fault diagnosis of rolling bearings based on EEMD permutation entropy and PCA-GK clustering", Journal of Henan University of Science and Technology(Natural Science), Vol 38, 2017, pp. 17-24.

[10] Case Western Reserve University. The case western reserve university bearing date center website bearing data center test seeded fault test data [EB/OL].(2016-05-04).