Application of a new clustering algorithm to analyze FT-IR spectrum data of lubricating oils

Jing Li, Hongxiang Tian, Tingfeng Ming, Yunling Sun and Shuai Zhang

College of Power Engineering, Naval University of Engineering.

Email: leejing0820@163.com

Abstract. Aiming at monitoring contaminated oil samples of marine engines, the FT-IR spectroscopy is used to analysis the oil samples. Under the condition of laboratory, 18 oil samples with different concentrations and different types of contaminants were measured. The types of contaminants were water, fuel dilution, ethylene glycol and oxidation. The FT-IR spectral data of oil samples is obtained. Firstly, the original FT-IR spectral data are pre-processed by the baseline correction and the data normalization. Then, the dimensions of FT-IR spectral data pre-processed are reduced by the principal component analysis (PCA) and the four different kinds of contaminant oil samples are shown by figures. Lastly, the low dimensional data are clustered by the self-organizing feature map network and the clustered results which stand for different kinds of contaminant oil samples are demonstrated in the digital form. The results showed that the accuracy of oxidation samples clustering reached 100%, the accuracy of contaminant water samples clustering reached 83%, and the accuracy of fuel or ethylene glycol contaminant samples clustering were unsatisfactory.

1. Introduction
The lubricating oil of marine diesel engine is an important part to ensure the normal running state of the diesel engine. The monitoring of lubricating oil of diesel engine needs to be sustained for a long time, which is of great significance to the safety of the ship. Some important equipment needs to accurately identify the contaminant type and status of the lubricating oil. In recent years, infrared spectroscopy analysis technology is becoming an important means of lubricating oil analysis for Marine diesel engines. For example, Shi Xin-fa et al. [1] studied 32 oil samples and extracted the information using principal component analysis. Wei Hai-jun et al. [2] established a mathematical model for the fault diagnosis of lubricating oil state based on spectral analysis by analyzing the long-term spectrum tracking of the lubricating oil of mechanical equipment. A rapid detection method of water contamination was developed by Zhang Yu [3] by extracting of Model input variable using principal component analysis (PCA) and continuous projection algorithm (SPA) method on visible - near infrared data. PCA was used to analysis spectrometric spectrometric data of sixty-nine oil samples by Liu tao [4] and clustering those data according to elements and oil samples separately. Max EPD was used to estimate the probability density distribution of oil monitoring data by Huo hua [5].
2. Data sources and algorithms

2.1. Data collection
The spectra were recorded using an Agilent Cary 5000 infrared spectrometer equipped with DialPath, which is configured with ZeSe type spectrum, and the spectral range was 4000 cm\(^{-1}\)-650 cm\(^{-1}\). The quantitative relationship between the characteristic value of the RT-IR spectral data measured and the concentration of contaminants can be established.

2.2. Samples and data
The experimental samples were divided into 5 groups, and different contaminants were mixed into the new oil as the lubricating oil samples. The four kinds of contaminants were water, oxidation, fuel dilution and ethylene glycol. Table 1 shows the configured samples number and corresponding contaminant state.

| Serial number | Component          | Configuration                  |
|---------------|--------------------|--------------------------------|
| 1-4           | Water              | Mix proportion:0.11%, 0.22%, 0.44%, 0.88%, 0% |
| 6-10          | Oxidation          | Heating time:200h, 299h, 323h, 371h, 395h |
| 11-14         | Diesel             | Mix proportion:1.5%, 3%, 6%, 12% |
| 16-18         | Ethylene Glycol    | Mix proportion:0.1%, 0.2%, 0.4% |
| 5, 15         | New oil            | 0                              |

2.3. Algorithm
After the spectral data pre-processing, the PCA method was adopted to reduce the dimensions, and then the self-organizing feature map network (SOM) algorithm was used for clustering. The detailed analysis method is as follows:

2.3.1. Sample. The spectra of 18 lubricating oil samples were analyzed by FT-IR spectrometer, and the absorbance curve was obtained. The horizontal coordinate was the wavenumber (cm\(^{-1}\)), and the interval scale was obtained.

2.3.2. Set up the sample data set. The spectral data was converted into ASCII code, and each sample curve sampled 900 points from 648 wavenumber to 3999 wavenumber, and obtained the sample data set \(S_{n \times d}(n=18,d=900)\). Where, \(n\) represents the number of samples, and \(d\) represents the sampling points of 900 equal interval wavenumber corresponding to each sample. The spectral characteristic band parameters were selected and the analysis matrix \(D\) was established.

2.3.3. Data pre-processing. The spectra were normalized and the characteristic bands were selected.

2.3.4. Reduction of data dimension using PCA. The specific algorithm is as follows:
   a) the column-wise mean of the matrix \(D_{ij}\) was calculated according to \(\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_i\);
   b) the data were centered, \(Z = D - 1 \cdot \mu^T\);
   c) calculation of the covariance matrix, \(\Sigma = \frac{1}{n} (Z^T Z)\);
   d) calculation of the characteristic values, \((\lambda_1, \lambda_2, \cdots, \lambda_d)\)=characteristic value (\(\Sigma\))
   e) calculate the eigenvector, \(U = (u_1 u_2 \cdots u_d)\)
   f) the proportion of the total variance, \(f(r) = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i} (r = 1, 2, \cdots, d)\)
   g) choose the minimum \(r\) of dimension, so that \(f(r) \geq \alpha\)
   h) after the dimension reduction, \(U_r = (u_1 u_2 \cdots u_r)\), and \(A = \{a_1| a_i = U_i^T x_i, i = 1, \cdots, n\}\)
2.3.5. **SOM algorithm.** The self-organizing feature map network generally has two layers, the input layer and the competition layer, where the competition layer is the core layer. Network to simulate the nerve cells in the brain response to external stimuli, by input samples repeatedly unsupervised learning, dissolves in the various characteristics of the input mode of connection weights, the implementation of a specific model for certain areas of the neurons in input response function, at the same time has a strong anti-interference ability.

Figure 1 shows the architecture of the network we selected. We have 13 input variables (The first 13 principal components were obtained through PCA analysis), and we are using 9 neurons. The feature map is 3×3, and it uses a hexagonal arrangement of neurons. This means that each internal neuron will have six neighbors.

Eq. (1) [5] is the SOM learning rule.

\[
w_i(q) = w_i(q - 1) + \alpha(p(q) - w_i(q - 1)) = (1 - \alpha)w_i(q - 1) + \alpha p(q)
\]

where \(i^*\) is the index of the winning neuron, and

\[
N_i(d) = \{j, \; d_{ij} \leq d\}
\]

defines the neuron neighborhood.

![Figure 1. SOM Network Architecture.](image)

### 3. Results and discussion

3.1. **Selection of characteristic bands**

The original data set is a set of infrared spectral curves obtained from the experiment of lubricating oil of different states. From 648 waves to 3999 wavenumbers, 900 points are sampled at intervals and an analytical matrix \(D_{ij}\) is obtained.

3.2. **PCA**

The training data set is used as the sample data (number 1-18) as the analysis object, and the principal component analysis is carried out. The main function of PCA is to reduce the data set and replace the high-dimensional data set with the new low-dimensional data set, and minimize the loss of information. The original data set is 18*286 dimensions, and the new matrix is obtained after analysis, and its eigenvalue and variation contribution are shown in Table 2. It can be seen that the first eight principal components in the new collection contain 99.85% of the original data.

The main components of the main components were scored in two dimensions and three dimensions. Figure 2 shows a 2D score plot of the second and fourth principal components. Figure 3 is the first, second and third component scores of the main components. It can be clearly seen that the training data set data can be divided into four categories according to the contaminant source.

3.3. **SOM**

For our trained SOM, for less than 25% of all input vectors, the winning neuron and the next closest neuron were not adjacent to each other. Figure 4 shows the Labeled SOM, to get more insight into how the SOM has clustered the data. In Figure 5, it is can be seen that 5 of the 19 oil samples are clustered into the first class, and 6 of the 19 oil samples are clustered into the second class, and the remaining oil samples are relatively scattered. The five oil samples in the first class are oxidized oil...
samples, and the clustering accuracy of the oxidation samples is 100%. The six oil samples in the second class were 5 water contaminant samples and a fuel oil sample, with a clustering accuracy of 83%. The more dispersed oil samples are from fuel and ethylene glycol contaminants.

Table 2. Principal component score sheet.

| Pcl | score | latent | tsquarel |
|-----|-------|--------|----------|
| 1   | 169.3 | 59.20  | 59.20    |
| 2   | 56.9  | 19.89  | 79.09    |
| 3   | 41.3  | 14.44  | 93.53    |
| 4   | 9.4   | 3.30   | 96.83    |
| 5   | 3.4   | 1.18   | 98.01    |
| 6   | 3.0   | 1.05   | 99.06    |
| 7   | 1.9   | 0.67   | 99.73    |
| 8   | 0.3   | 0.12   | 99.85    |

Figure 2. The Second - Fourth Principal Component Score Chart.

Figure 3. The First-Three Principal Component Scores Chart.
The results are not discussed by relating the findings to previous studies.

4. Conclusions
It can effectively solve the problem of reducing the contaminant data of marine diesel engines using PCA. The characteristic band values associated with water, oxidation, fuel and ethylene glycol contaminant were obtained. The oil samples are visually classified by different types of contaminants. The classification depends on the subjective feeling of the individual, and the individual difference affects the final result.

Using the SOM algorithm to analyze the low dimensional data, the clustering of different oil samples can be further quantified. The results showed that the accuracy of the oxidized contaminant cluster was 100%, the accuracy of water contaminant clustering was 83%, and the accuracy of contaminant such as fuel and ethylene glycol was not satisfactory.

References
[1] Shi X F, Liu D F, Zhou Z C 2015 Engine Engineering 34(5) 81
[2] Wei H J, Wang H Z, Sun P T 2005 Spectroscopy and Spectral Analysis 25(7) 1125
[3] Zhang Y, Jiang L L, Wu D, Tan L H 2010 Spectroscopy and Spectral Analysis 30(8) 2111
[4] Liu T, Tian H X, Gong W Y 2010 Spectroscopy and Spectral Analysis 30(3) 779
[5] Huo H, Li Z G 2005 JOURNAL OF SHANGHAI JIAOTONG UNIVERSITY 39(1) 95