# Clustering of Infrared Spectra of Lubricating Base Oils Using Adaptive Resonance Theory

X. Z. Wang* and B. H. Chen

Department of Chemical Engineering, The University of Leeds, Leeds LS2 9JT, U.K.

Received November 24, 1997

Adaptive resonance theory (ART2) is applied to the classification of lubricating base oils on the basis of their infrared spectra. Fifty-nine data samples are used which were collected from 12 refineries representing eight crude oil origins. The ART2 neural network is an unsupervised machine learning approach which is different from supervised learning, such as back-propagation neural networks, in that it determines both the number of classes and the assignments of data samples. The ART2 groups the 59 data samples into seven classes. Five of the seven classes are found to perfectly match the crude oil origins. Two of the seven classes (eight samples in total) are combined to form a single class. The class assignment of one data sample (sample 35) does not match the crude oil origin but is consistent with the prior observation of its spectrum. The work demonstrates that ART2 represents a useful alternative tool for infrared spectra interpretation.

## 1. INTRODUCTION

The fundamental factor determining the application performance of lubricating oils is known to be the structural composition. A great deal of effort has been devoted to the characterization of petroleum samples on the basis of gas chromatograms, infrared spectra, fluorescence spectra, and other laboratory data.[1−4] Some of these techniques have been refined to the point where they can be applied to the typing of unknown petroleum samples. One of the main purposes of the research has been to establish the linear and nonlinear relationships between application performance and structural compositions, which clearly depend on the proper interpretation of the data collected.

Since the 1960s, with the development of computational chemistry, various factor analysis and pattern recognition technologies have been applied to analytical chemistry to process high-dimensional data.[5−12] In recent years, there has been a growing interest in applying neural networks to the classification of IR spectra.[2−4,13−20] Most studies have focused on back-propagation neural networks (BPNN). BPNN uses a set of data with known classifications as training patterns. The training (or learning) process searches for the optimum weights connecting the inputs (here attributes) and outputs (here classification schemes). Once trained, a BPNN can be used to predict the classification of unclassified data. BPNN has been proved to be an effective tool in the interpretation of infrared spectra. However, it still has some limitations.[18−20] The major limitation is that it is an supervised learning approach, which means that it always requires a set of input−output pairs (known data). Through empirical learning, it captures the features of training data patterns rather than the features of unclassified patterns. It has been well recognized that BPNN has the potential "extrapolation error".[21]

Data encountered can be in the following three forms.

(1) Part of the database is known, i.e., the number and descriptions of classes as well as the assignments of individual data patterns are known. The task is to assign the unknown data patterns to the established classes.

(2) Only the number and descriptions of classes are known. The task is to assign all data patterns to the known classes.

(3) Both the number and descriptions of classes are not known. The task is to determine the number and descriptions of classes as well as the assignments of data patterns.

BPNN is a powerful tool in dealing with the first type of data. However, clearly it is not suitable for the last two situations, since training data patterns are not available. For the last two situations, an *unsupervised* learning approach is needed. Schuur et al.,[22] Novic and Zupan,[20] and Zupan et al.[18,19] have applied the self-organizing Kohonen and counter-propagation networks to the infrared spectra analysis. The Kohonen network is an unsupervised approach, which is used to map all the spectra into groups. The preliminary grouping is used for the separation of spectra into the training and the test sets for further use by a counter-propagation network. The combination of both unsupervised and supervised learning gave results that are superior to those from a random selection of training data sets.

The present work has focused on the application of an *unsupervised* neural network—the adaptive resonance theory (ART)—to the classification of infrared spectra of lubricating base oils. ART is able to determine both the number of classes and the assignments of individual data patterns through learning. The classification scheme will group objects (or data patterns) into a number of classes such that objects within classes are similar in some respect and unlike those from other classes. The ART algorithm will be briefly introduced in the second section. The application of the approach to the classification of the 59 infrared spectra of lubricating base oils will be presented in the third section.

* Corresponding author. Tel.: +44 113 233 2427. Fax: +44 113 233 2405. E-mail: che6xzw@sun.leeds.ac.uk.

S0095-2338(97)00434-4 CCC: $15.00 
Published on Web 04/17/1998

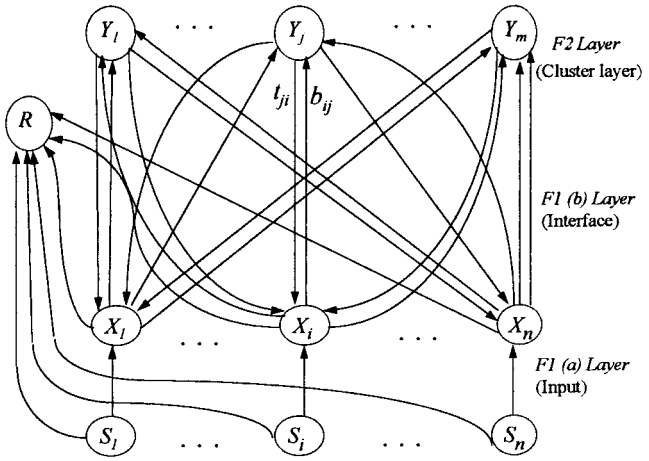


**Figure 1.** General architecture of ART.

The current approach is compared with others also in the third section, which is followed by conclusions.

## 2. ADAPTIVE RESONANCE THEORY

Adaptive resonance theory (ART) was developed by Grossberg[23,24] as a clustering-based, autonomous learning model. ART has been instantiated in a series of neural network models referred to as ART1, ART2, and ART3.[25−28] ART1 is designed for clustering binary vectors, while ART2 is for continuous-valued vectors.

A general architecture of ART neural network is shown in Figure 1. An ART network consists of three groups of neurons, the input processing layer ($F_1$ in Figure 1), the cluster layer ($F_2$ in Figure 1), and a mechanism to control the degree of similarity of patterns placed on the same cluster (a reset mechanism). The input layer can be considered to consist of two parts: the input portion and the interface portion. Some processing may occur in the input portion and the interface portion (especially in ART2). The interface portion combines signals from the input signals to the weight vector for the cluster unit that has been selected as a candidate for learning. The input portion and the interface portion are denoted as $F_1$ (a) and $F_1$ (b).

An input data pattern is presented to the network, and after some specialized processing by the network, the input pattern is compared to each of the existing cluster prototypes in the cluster layer ($F_2$). The "winner" in the cluster layer is the prototype most similar to the input. Whether or not a cluster unit is allowed to learn an input data pattern depends on how similar its top-down weight vector is to the input vector. The decision is made by the reset unit, based on signals it receives from the input (a) and the interface (b) portions of the $F_1$ layer. If the similarity between the "winner" and the input exceeds a predetermined value (the vigilance parameter $\rho$ in ART terms), learning is enabled, and the "winner" is modified slightly to more closely reflect the input data pattern. If the similarity between the "winner" and the input is less than required by the vigilance parameters, a new cluster unit is initialized in the top layer, and learning is enabled to create a prototype similar to the input. It is clear that both the number and descriptions of classes are continuously updated during learning. This is different from the Kohonen network. Though the descriptions of classes (called neurons in a Kohonen network) in a Kohonen network also continuously update during learning, the number of classes needs to be determined before learning starts.

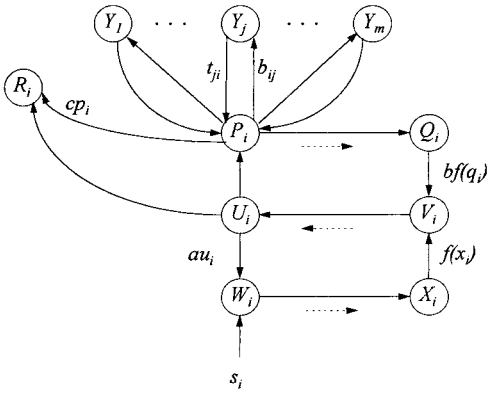

**Figure 2.** Typical ART2 architecture.

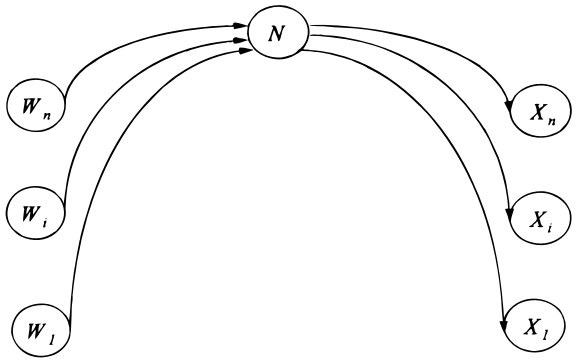

**Figure 3.** Details of connections from *W* to *X* units, showing the supplemental unit *N* to perform normalization.

To control the similarity of patterns placed on the same cluster, there are two sets of connections (each with its own weights) between each unit in the interface portion of the input layer and each cluster unit. The $F_1$ (b) layer is connected to the $F_2$ layer by bottom-up weights; the bottom-up weight on the connection from the $i$th $F_1$ unit to the $j$th $F_2$ unit is designated $b_{ij}$. The $F_2$ layer is connected to the $F_1$ (b) layer by top-down weights; the top-down weight on the connection from the $j$th $F_2$ unit to the $i$th $F_1$ unit is designated $t_{ji}$.

ART2 is designed to process continuous-valued data patterns. A typical ART2 architecture[25] is illustrated in Figure 2. It has a very sophisticated input data processing layer ($F_1$ ) consisting of six types of units (the *W*, *X*, *U*, *V*, *P*, and *Q* units). There are *n* units of each of these types (where *n* is the dimension of an input pattern). Only one unit of each type is shown in Figure 2. A supplemental unit between the *W* and *X* units receives signals from all of the *W* units, computes the norm of the vector **w**, and sends this (inhibitory) signal to each of the *X* units. Each of these also receives an excitatory signal from the corresponding *W* unit. Details of this portion of the net are shown in Figure 3. A similar supplemental unit performs the same role between the *P* and *Q* units, and another does so between the *V* and *U* units. Each *X* unit is connected to the corresponding *V* unit, and each *Q* unit is connected to the corresponding *V* unit also.

The symbols on the connection paths between the various units in the $F_1$ layer in Figure 2 indicate the transformation that occurs to the signals as it passes from one type of unit

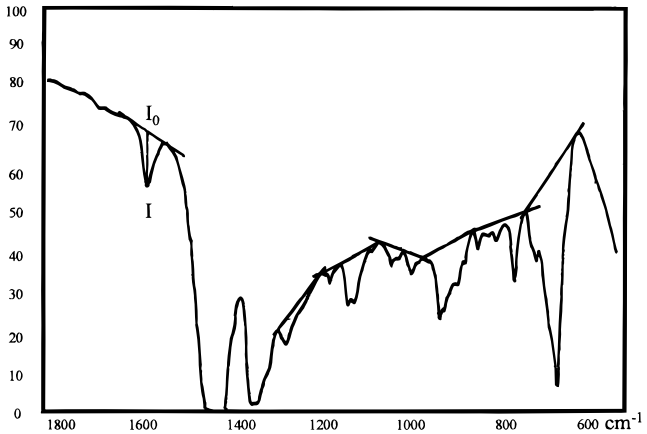


**Figure 4.** Baseline approach to find $I_0$ and $I$ in eq 1.

to the next; they do not indicate multiplication by the given quantity. However, the connections between units $P_i$ (of the $F_i$ layer) and $Y_i$ (of the $F_2$ layer) do show the weights that multiply the signal transmitted over those paths. The activation of the winning $F_2$ unit is $d$, where $0 < d < 1$. The symbol $\rightarrow$ indicates normalization; i.e., the vector **q** of activations of $Q$ units is just the vector **p** of activations of the $P$ units, normalized to approximately unit length.

The $U$ units perform the role of an input phase of the $F_1$ layer. The $P$ units play the role of the interface of the $F_1$ layer. Units $X_i$ and $Q_i$ apply an activation function to their net input; this function suppresses any components of the vectors of activations at those levels that fall below the user-selected value $\theta$. The connection paths from $W$ to $U$ and from $Q$ to $V$ have fixed weights $a$ and $b$, respectively.

The most complete description of the ART2 network can be found.[25] Although written for ART1, the discussions by Caudil[29] and Wasserman[30] provide a useful but less technical description of many of the principles employed in ART2.

## 3. AUTOMATIC CLASSIFICATION OF LUBRICATING BASE OILS

**The Data Sets.** The data used in this study are taken from Zhou,[31] who analyzed 59 samples of lubricating base oils collected from 12 refineries using an IR-27G infrared spectrometer. Zhou[31,32] compared various parameters proposed in the literature to capture the features of a spectrum, such as the peak heights and areas,[5,33,6,11] and found that the absorption strength of an infrared band defined by eq 1 can best represent the spectra of lubricating base oils.

$$\epsilon^i_{\max} = \frac{lg(I_0/I)_i}{L} \tag{1}$$

where $\epsilon^i_{\max}$ is the absorption strength of the $i$th infrared band, $I_0$ and $I$ are to be estimated using the baseline approach, as shown in Figure 4, and $L$ is path length of the sodium chloride cells, here 0.206 mm.

Twenty-one peaks were used to characterize a spectrum: 1604 ± 1.8, 1304 ± 2.5, 1209 ± 2.2, 1168 ± 1.3, 1153 ± 1.9, 1119 ± 2.0, 1075 ± 1.8, 1061 ± 1.2, 1031 ± 1.1, 973 ± 1.2, 967 ± 4.2, 955 ± 1.8, 932 ± 1.1, 916 ± 1.2, 886 ± 1.6, 868 ± 2.6, 847 ± 1.7, 810 ± 0.8, 761 ± 1.6, 731 ± 2.7, and 720 ± 1.1 cm$^{-1}$. The advantage of the approach used by Zhou[31] is that the bands around the fingerprint

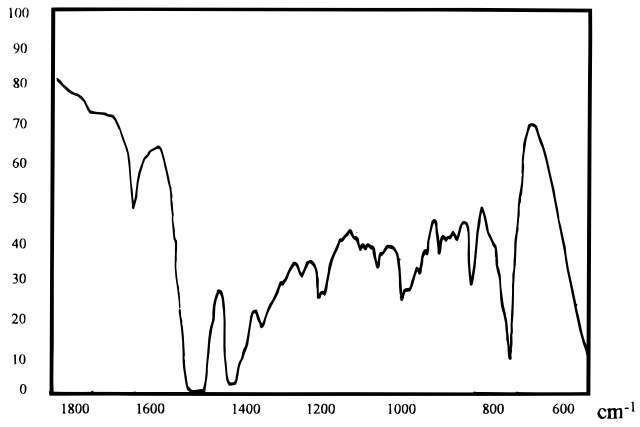


**Figure 5.** Spectrum of sample 41 (oil source, Yumen; refinery, R-YM).

**Table 1.** Source of Data

| oil source | no. of samples | refineries, sample numbers |
|---|---|---|
| Daqing | 27 | R-7, 6−11, 19, 21, 23, 26, 27<br>R-SH, 1−5, 22, 24<br>R-5, 17, 18, 20<br>R-MM, 12−15<br>R-DFH, 16 |
| Xinjiang | 13 | R-DSZ, 28−31<br>R-LL, 32−34, 37−39<br>R-XJ, 36, 40<br>R-KLM, 35[a] |
| Dagang | 6 | R-DG, 44−47, 50−51 |
| Renqiu | 4 | R-DL, 56−59[b] |
| Linshang | 2 | R-JN, 48−49 |
| Yumen | 3 | R-YM, 41−43 |
| Nanning | 2 | R-JM, 52−53 |
| Nanchong | 2 | R-NC, 54−55 |

[a] Sample 35 is noticeably close to those from Dagang in spectrum though it was collected from the refinery R-KLM, which used Xinjiang oil. [b] The four samples 56−59 are produced by refining a mixture of both Renqiu and Daqing oils.

(1200−720 cm$^{-1}$) are very well distinguished, which can be seen from Figures 4 and 5. The 59 data sets are summarized in Tables 1 and 2. Prior observations were recorded for the spectra of samples 35 and 56−59.[31] Sample 35 was noticeably close to those from Dagang oil in spectra, although it was collected from the refinery R-KLM, which used Xingjiang oil. Samples 56−59 were collected from the refinery R-DL, which used a mixture of Renqiu and Daqing oils.

**Classification Results.** The task becomes a problem of 59 samples, each having 21 attributes, as shown in Table 2. The ART2 neural network generates a number of classification schemes according to the selection of the vigilance value. The final decision is to be made by domain experts after careful study of various classification schemes. The best classification for the current problem is shown in Table 3.

**Analysis of the Classification Results.** The best classification is obtained when the vigilance parameter $\rho$ = 0.9955 and is summarized in Table 3. Basically, the 59 data samples are grouped into seven classes. The classification is surprisingly good. Classes I, II III, IV, and VII are found to perfectly match the crude oil origins. Classes V and VI provide the basis for making the final classification decision. When inspecting the results, it is helpful to always make reference to the original data summarized in Table 1.

**Table 2.** Absorption Strength, Defined by Eq 1 for the 21 Bands

| | 1604 $cm^{-1}$ | 1304 $cm^{-1}$ | 1207 $cm^{-1}$ | 1169 $cm^{-1}$ | 1153 $cm^{-1}$ | 1119 $cm^{-1}$ | 1071 $cm^{-1}$ | 1059 $cm^{-1}$ | 1031 $cm^{-1}$ | 972 $cm^{-1}$ | 967 $cm^{-1}$ | 953 $cm^{-1}$ | 932 $cm^{-1}$ | 916 $cm^{-1}$ | 886 $cm^{-1}$ | 868 $cm^{-1}$ | 847 $cm^{-1}$ | 810 $cm^{-1}$ | 763 $cm^{-1}$ | 732 $cm^{-1}$ | 720 $cm^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.459 | 0.966 | 0.085 | 0.5 | 0.564 | 0.033 | 0.249 | 0.176 | 0.209 | 0.0 | 0.821 | 0.0 | 0.333 | 0.215 | 0.530 | 0.214 | 0.184 | 0.554 | 0.913 | 2.679 | 5.690 |
| 2 | 0.580 | 0.946 | 0.196 | 0.430 | 0.516 | 0.038 | 0.282 | 0.181 | 0.275 | 0.0 | 1.150 | 0.0 | 0.449 | 0.133 | 0.427 | 0.268 | 0.073 | 0.705 | 0.912 | 2.397 | 5.286 |
| 3 | 0.662 | 0.936 | 0.176 | 0.393 | 0.412 | 0.030 | 0.293 | 0.221 | 0.282 | 0.0 | 0.995 | 0.0 | 0.293 | 0.108 | 0.533 | 0.369 | 0.116 | 0.686 | 0.969 | 2.425 | 5.709 |
| 4 | 0.742 | 1.018 | 0.109 | 0.317 | 0.424 | 0.022 | 0.341 | 0.216 | 0.197 | 0.0 | 0.905 | 0.0 | 0.206 | 0.076 | 0.492 | 0.390 | 0.125 | 0.641 | 0.999 | 2.426 | 6.249 |
| 5 | 0.675 | 0.896 | 0.016 | 0.289 | 0.401 | 0.010 | 0.348 | 0.250 | 0.275 | 0.0 | 0.999 | 0.0 | 0.159 | 0.032 | 0.420 | 0.307 | 0.11 | 0.475 | 0.911 | 2.911 | 6.509 |
| 6 | 0.449 | 0.960 | 0.064 | 0.557 | 0.621 | 0.019 | 0.266 | 0.199 | 0.136 | 0.0 | 1.224 | 0.0 | 0.340 | 0.212 | 0.455 | 0.137 | 0.108 | 0.472 | 0.855 | 2.676 | 5.316 |
| 7 | 0.405 | 0.855 | 0.147 | 0.550 | 0.614 | 0.027 | 0.218 | 0.148 | 0.178 | 0.0 | 1.184 | 0.0 | 0.400 | 0.197 | 0.379 | 0.118 | 0.132 | 0.442 | 0.873 | 2.499 | 5.311 |
| 8 | 0.365 | 0.851 | 0.163 | 0.494 | 0.584 | 0.034 | 0.335 | 0.150 | 0.162 | 0.0 | 1.204 | 0.0 | 0.435 | 0.226 | 0.306 | 0.107 | 0.146 | 0.436 | 0.791 | 2.514 | 5.467 |
| 9 | 0.627 | 0.875 | 0.19 | 0.404 | 0.489 | 0.033 | 0.324 | 0.237 | 0.322 | 0.0 | 1.054 | 0.0 | 0.327 | 0.111 | 0.433 | 0.210 | 0.134 | 0.582 | 0.902 | 2.393 | 5.571 |
| 10 | 0.694 | 0.944 | 0.174 | 0.381 | 0.468 | 0.020 | 0.342 | 0.223 | 0.189 | 0.0 | 0.998 | 0.0 | 0.471 | 0.087 | 0.438 | 0.287 | 0.066 | 0.572 | 0.934 | 2.7 | 5.863 |
| 11 | 0.637 | 0.917 | 0.062 | 0.259 | 0.401 | 0.010 | 0.217 | 0.234 | 0.287 | 0.0 | 0.953 | 0.0 | 0.177 | 0.048 | 0.403 | 0.276 | 0.105 | 0.430 | 0.945 | 2.572 | 6.204 |
| 12 | 0.312 | 0.844 | 0.193 | 0.501 | 0.591 | 0.029 | 0.226 | 0.157 | 0.150 | 0.0 | 1.056 | 0.0 | 0.368 | 0.126 | 0.374 | 0.186 | 0.135 | 0.544 | 0.851 | 2.532 | 5.743 |
| 13 | 0.544 | 0.886 | 0.225 | 0.413 | 0.413 | 0.034 | 0.292 | 0.220 | 0.346 | 0.0 | 1.105 | 0.0 | 0.482 | 0.238 | 0.285 | 0.083 | 0.155 | 0.372 | 0.742 | 2.596 | 5.315 |
| 14 | 0.845 | 0.930 | 0.157 | 0.371 | 0.452 | 0.016 | 0.383 | 0.266 | 0.387 | 0.0 | 0.923 | 0.0 | 0.234 | 0.070 | 0.481 | 0.382 | 0.115 | 0.680 | 1.008 | 2.604 | 5.879 |
| 15 | 1.055 | 0.989 | 0.079 | 0.291 | 0.405 | 0.042 | 0.364 | 0.288 | 0.379 | 0.0 | 0.920 | 0.0 | 0.188 | 0.041 | 0.536 | 0.477 | 0.177 | 0.712 | 1.046 | 2.470 | 5.063 |
| 16 | 0.428 | 0.829 | 0.207 | 0.523 | 0.581 | 0.050 | 0.261 | 0.135 | 0.166 | 0.0 | 0.828 | 0.0 | 0.451 | 0.235 | 0.361 | 0.157 | 0.168 | 0.465 | 0.824 | 2.650 | 5.430 |
| 17 | 0.468 | 0.838 | 0.206 | 0.501 | 0.575 | 0.082 | 0.275 | 0.147 | 0.235 | 0.0 | 1.067 | 0.0 | 0.430 | 0.200 | 0.430 | 0.156 | 0.124 | 0.562 | 0.869 | 2.493 | 5.307 |
| 18 | 0.555 | 0.861 | 0.210 | 0.426 | 0.528 | 0.058 | 0.267 | 0.189 | 0.233 | 0.0 | 1.037 | 0.0 | 0.426 | 0.164 | 0.414 | 0.203 | 0.150 | 0.668 | 0.875 | 2.412 | 5.239 |
| 19 | 0.417 | 0.853 | 0.234 | 0.453 | 0.538 | 0.050 | 0.288 | 0.128 | 0.256 | 0.0 | 1.182 | 0.0 | 0.415 | 0.154 | 0.356 | 0.142 | 0.144 | 0.460 | 0.797 | 2.464 | 5.287 |
| 20 | 0.376 | 0.780 | 0.206 | 0.520 | 0.603 | 0.053 | 0.226 | 0.149 | 0.162 | 0.0 | 1.140 | 0.0 | 0.410 | 0.144 | 0.402 | 0.123 | 0.120 | 0.419 | 0.813 | 2.378 | 5.235 |
| 21 | 0.354 | 0.902 | 0.268 | 0.488 | 0.568 | 0.050 | 0.268 | 0.152 | 0.228 | 0.0 | 0.985 | 0.0 | 0.504 | 0.198 | 0.274 | 0.130 | 0.159 | 0.420 | 0.714 | 2.435 | 5.205 |
| 22 | 0.398 | 0.855 | 0.222 | 0.469 | 0.548 | 0.043 | 0.278 | 0.174 | 0.196 | 0.0 | 0.945 | 0.0 | 0.401 | 0.192 | 0.322 | 0.165 | 0.166 | 0.448 | 0.788 | 2.464 | 5.350 |
| 23 | 0.663 | 0.904 | 0.202 | 0.385 | 0.476 | 0.036 | 0.346 | 0.210 | 0.265 | 0.0 | 1.085 | 0.0 | 0.338 | 0.096 | 0.436 | 0.275 | 0.139 | 0.630 | 0.897 | 2.494 | 5.388 |
| 24 | 0.373 | 0.881 | 0.246 | 0.507 | 0.590 | 0.065 | 0.284 | 0.117 | 0.156 | 0.0 | 1.224 | 0.0 | 0.506 | 0.196 | 0.306 | 0.127 | 0.138 | 0.428 | 0.775 | 2.521 | 5.287 |
| 25 | 0.391 | 0.875 | 0.232 | 0.481 | 0.574 | 0.058 | 0.288 | 0.118 | 0.223 | 0.0 | 1.191 | 0.0 | 0.491 | 0.177 | 0.283 | 0.183 | 0.158 | 0.434 | 0.768 | 2.443 | 5.288 |
| 26 | 0.528 | 1.083 | 0.167 | 0.451 | 0.581 | 0.037 | 0.405 | 0.263 | 0.307 | 0.0 | 0.613 | 0.0 | 0.478 | 0.223 | 0.422 | 0.123 | 0.140 | 0.402 | 0.896 | 2.438 | 5.582 |
| 27 | 0.535 | 0.861 | 0.204 | 0.548 | 0.586 | 0.044 | 0.238 | 0.170 | 0.205 | 0.0 | 0.703 | 0.0 | 0.458 | 0.292 | 0.453 | 0.249 | 0.170 | 0.606 | 0.883 | 2.506 | 5.218 |
| 28 | 0.233 | 0.558 | 0.298 | 1.071 | 0.968 | 0.025 | 0.133 | 0.166 | 0.191 | 1.230 | 0.0 | 0.861 | 0.762 | 0.530 | 0.268 | 0.058 | 0.319 | 0.444 | 0.670 | 2.408 | 3.239 |
| 29 | 0.184 | 0.496 | 0.349 | 0.944 | 0.921 | 0.018 | 1.170 | 0.205 | 0.192 | 1.541 | 0.0 | 1.090 | 0.823 | 0.558 | 0.279 | 0.071 | 0.362 | 0.443 | 0.574 | 2.287 | 2.958 |
| 30 | 0.232 | 0.429 | 0.567 | 0.886 | 0.864 | 0.014 | 0.253 | 0.398 | 0.336 | 1.837 | 0.0 | 1.285 | 0.963 | 0.531 | 0.258 | 0.120 | 0.445 | 0.499 | 0.559 | 2.431 | 2.900 |
| 31 | 0.247 | 0.553 | 0.509 | 0.777 | 0.805 | 0.0 | 0.207 | 0.364 | 0.253 | 1.594 | 0.0 | 1.085 | 0.821 | 0.436 | 0.332 | 0.162 | 0.399 | 0.451 | 0.571 | 2.758 | 3.556 |
| 32 | 0.432 | 0.550 | 0.264 | 1.010 | 0.962 | 0.097 | 0.129 | 0.191 | 0.264 | 1.182 | 0.0 | 0.898 | 0.753 | 0.488 | 0.272 | 0.192 | 0.323 | 0.563 | 0.698 | 2.456 | 3.259 |
| 33 | 0.498 | 0.568 | 0.276 | 0.865 | 0.864 | 0.097 | 0.218 | 0.209 | 0.320 | 1.279 | 0.0 | 0.894 | 0.755 | 0.436 | 0.287 | 0.228 | 0.328 | 0.740 | 0.658 | 2.449 | 3.555 |
| 34 | 0.364 | 0.572 | 0.430 | 0.897 | 0.895 | 0.067 | 0.187 | 0.274 | 0.282 | 1.423 | 0.0 | 1.077 | 0.843 | 0.477 | 0.267 | 0.189 | 0.374 | 0.522 | 0.609 | 2.544 | 3.283 |
| 35 | 0.442 | 0.414 | 0.157 | 1.075 | 0.883 | 0.074 | 0.097 | 0.071 | 0.157 | 0.969 | 0.0 | 0.712 | 0.655 | 0.551 | 0.245 | 0.187 | 0.371 | 0.652 | 0.765 | 1.704 | 1.993 |
| 36 | 0.300 | 0.517 | 0.272 | 1.052 | 1.013 | 0.066 | 0.138 | 0.144 | 0.141 | 1.324 | 0.0 | 0.945 | 0.786 | 0.535 | 0.254 | 0.155 | 0.341 | 0.503 | 0.659 | 2.401 | 3.237 |
| 37 | 0.258 | 0.584 | 0.275 | 1.068 | 1.003 | 0.071 | 0.144 | 0.158 | 0.155 | 1.190 | 0.0 | 0.943 | 0.802 | 0.515 | 0.209 | 0.097 | 0.333 | 0.420 | 0.662 | 2.436 | 3.245 |
| 38 | 0.196 | 0.603 | 0.314 | 1.028 | 1.093 | 0.078 | 0.133 | 0.153 | 0.126 | 1.228 | 0.0 | 1.015 | 0.824 | 0.577 | 0.220 | 0.075 | 0.336 | 0.375 | 0.635 | 2.550 | 3.260 |
| 39 | 0.304 | 0.545 | 0.267 | 1.021 | 1.060 | 0.066 | 0.124 | 0.148 | 0.116 | 0.908 | 0.0 | 0.943 | 0.792 | 0.503 | 0.259 | 0.146 | 0.342 | 0.505 | 0.670 | 2.394 | 3.178 |
| 40 | 0.234 | 0.480 | 0.408 | 0.965 | 0.921 | 0.0 | 0.155 | 0.199 | 0.214 | 1.605 | 0.0 | 1.346 | 0.889 | 0.544 | 0.227 | 0.072 | 0.393 | 0.455 | 0.571 | 2.221 | 2.715 |
| 41 | 0.697 | 0.649 | 0.202 | 0.797 | 0.763 | 0.090 | 0.205 | 0.168 | 0.361 | 1.036 | 0.0 | 0.906 | 0.630 | 0.394 | 0.413 | 0.245 | 0.309 | 1.013 | 0.691 | 2.083 | 3.857 |
| 42 | 0.860 | 0.562 | 0.319 | 0.733 | 0.770 | 0.243 | 0.253 | 0.161 | 0.459 | 1.456 | 0.0 | 1.180 | 0.800 | 0.453 | 0.416 | 0.338 | 0.480 | 1.259 | 0.677 | 2.064 | 3.232 |
| 43 | 1.223 | 0.588 | 0.424 | 0.523 | 0.659 | 0.127 | 0.255 | 0.218 | 0.435 | 1.343 | 0.0 | 0.986 | 0.661 | 0.369 | 0.508 | 0.533 | 0.531 | 1.315 | 0.778 | 2.214 | 3.460 |
| 44 | 0.702 | 0.423 | 0.179 | 0.890 | 0.816 | 0.0 | 0.175 | 0.179 | 0.265 | 0.805 | 0.0 | 0.679 | 0.598 | 0.420 | 0.364 | 0.349 | 0.404 | 1.199 | 1.142 | 1.852 | 2.131 |
| 45 | 0.652 | 0.468 | 0.177 | 0.852 | 0.826 | 0.068 | 0.176 | 0.155 | 0.253 | 0.910 | 0.0 | 0.752 | 0.633 | 0.424 | 0.353 | 0.332 | 0.375 | 1.207 | 1.069 | 1.849 | 2.306 |
| 46 | 0.593 | 0.495 | 0.265 | 0.873 | 0.829 | 0.0 | 0.166 | 0.151 | 0.256 | 1.046 | 0.0 | 0.829 | 0.724 | 0.446 | 0.312 | 0.286 | 0.302 | 1.220 | 0.919 | 1.888 | 2.526 |
| 47 | 0.568 | 0.537 | 0.247 | 0.822 | 0.819 | 0.103 | 0.151 | 0.111 | 0.245 | 1.016 | 0.0 | 0.844 | 0.741 | 0.470 | 0.277 | 0.237 | 0.280 | 1.225 | 0.853 | 1.916 | 2.607 |
| 48 | 0.451 | 0.594 | 0.252 | 0.742 | 0.757 | 0.084 | 0.244 | 0.134 | 0.263 | 1.158 | 0.0 | 0.784 | 0.642 | 0.327 | 0.266 | 0.13 | 0.211 | 0.799 | 0.799 | 2.202 | 4.456 |
| 49 | 0.370 | 0.534 | 0.500 | 0.661 | 0.778 | 0.329 | 0.434 | 0.0 | 0.493 | 2.002 | 0.0 | 1.136 | 0.766 | 0.408 | 0.181 | 0.0 | 0.358 | 0.682 | 0.656 | 2.135 | 4.055 |
| 50 | 0.569 | 0.470 | 0.290 | 0.760 | 0.808 | 0.172 | 0.171 | 0.079 | 0.323 | 1.422 | 0.0 | 1.063 | 0.804 | 0.480 | 0.266 | 0.280 | 0.360 | 1.340 | 0.748 | 1.838 | 2.514 |
| 51 | 0.537 | 0.523 | 0.384 | 0.790 | 0.817 | 0.207 | 0.204 | 0.086 | 0.350 | 1.507 | 0.0 | 1.110 | 0.844 | 0.486 | 0.265 | 0.255 | 0.388 | 1.342 | 0.651 | 1.797 | 2.531 |
| 52 | 0.553 | 0.716 | 0.238 | 0.852 | 0.827 | 0.095 | 0.238 | 0.214 | 0.242 | 1.045 | 0.0 | 0.821 | 0.642 | 0.425 | 0.271 | 0.093 | 0.168 | 0.681 | 0.771 | 2.485 | 4.490 |
| 53 | 0.568 | 0.575 | 0.474 | 0.759 | 0.811 | 0.266 | 0.380 | 0.265 | 0.435 | 1.711 | 0.0 | 1.176 | 0.797 | 0.461 | 0.218 | 0.189 | 0.313 | 0.817 | 0.623 | 2.226 | 3.819 |
| 54 | 1.056 | 0.785 | 0.118 | 0.470 | 0.523 | 0.0 | 0.278 | 0.201 | 0.446 | 0.0 | 0.443 | 0.0 | 0.390 | 0.171 | 0.647 | 0.751 | 0.280 | 1.292 | 0.633 | 2.692 | 4.990 |
| 55 | 0.988 | 0.821 | 0.154 | 0.451 | 0.496 | 0.015 | 0.273 | 0.184 | 0.362 | 0.0 | 0.470 | 0.0 | 0.394 | 0.149 | 0.603 | 0.612 | 0.220 | 1.056 | 1.136 | 2.346 | 5.131 |
| 56 | 0.432 | 0.600 | 0.242 | 1.006 | 0.968 | 0.315 | 0.195 | 0.071 | 0.316 | 1.112 | 0.0 | 1.134 | 0.710 | 0.508 | 0.207 | 0.041 | 0.219 | 0.882 | 0.689 | 2.323 | 3.871 |
| 57 | 0.332 | 0.584 | 0.300 | 0.969 | 0.994 | 0.537 | 0.212 | 0.095 | 0.395 | 1.504 | 0.0 | 1.564 | 0.822 | 0.581 | 0.178 | 0.0 | 0.344 | 0.816 | 0.603 | 2.015 | 3.636 |
| 58 | 0.493 | 0.545 | 0.379 | 0.866 | 0.870 | 0.413 | 0.314 | 0.218 | 0.390 | 1.580 | 0.0 | 1.436 | 0.807 | 0.547 | 0.196 | 0.153 | 0.365 | 0.821 | 0.666 | 2.226 | 3.819 |
| 59 | 0.840 | 0.584 | 0.268 | 0.601 | 0.698 | 0.202 | 0.245 | 0.195 | 0.306 | 1.043 | 0.0 | 0.961 | 0.645 | 0.380 | 0.268 | 0.348 | 0.356 | 0.770 | 0.740 | 2.546 | 4.276 |

**Table 3.** Classification Results When the Vigilance Parameter $\rho = 0.9955$

| class | no. of samples | samples |
|---|---|---|
| I | 27 | 1−27 |
| II | 12 | 28−34, 36−40 |
| III | 7 | 35, 44−47, 50, 51 |
| IV | 3 | 41−43 |
| V | 4 | 48, 52, 56, 59 |
| VI | 4 | 49, 53, 57, 58 |
| VII | 2 | 54, 55 |

Class I consists of 27 data samples from five refineries, R-7 (samples 6−11, 19, 21, 23, 26, and 27), R-SH (samples 1−5, 22, and 24), R-5 (samples 17, 18, and 20), R-MM (samples 12−15), and R-DFH (sample 16). Although the 27 samples are from five different refineries having varied processing strategies and technologies, which will, no doubt, have an influence on the properties of lubricating base oils, they represent the same crude oil origin, which is Daqing (see the second row of Table 1), and are more similar than those from a different crude oil origin.

Class II includes 12 data samples from four refineries, including refineries R-DSZ (samples 28−31), R-LL (samples 32−34, 37−39), R-XJ (samples 36 and 40). They all

represent the same crude oil origin, which is Xinjiang (see row 3 in Table 1). Specific attention is given to sample 35, collected from the refinery R-KLM, which is not classified to class II. Instead, it is assigned to class III. This is not beyond our expectation, since, as indicated earlier (the footnote of Table 1 about sample 35), sample 35 was collected from R-KLM, which used mixed crude oils of Xinjiang and Dagang. As will be discussed later, class III corresponds to Dagang oil. Therefore, we would regard the assignment of 35 as correct.

Class III consists of the seven data samples from refineries R-DG (samples 44−47, 50, 51) and R−KLM (sample 35). Samples 44−47 and 50, 51, though from two different refineries, represent the same crude oil origin, Dagang (see row 4 of Table 1). The assignment of sample 35 has been justified above. Therefore, class III is also a perfect classification.

Class IV includes all three samples (41−43) from the refinery R-YM, which have the same crude oil origin, Yumen (see row 7 in Table 1), so it is also a perfect classification.

Class VII consists of the only two samples (samples 54 and 55) collected from a refinery (R-NC) which used Nanchong crude oil (see the last row of Table 1). A major challenge to data clustering has been the identification of a small volume of data in a large database. When they are included in a large volumes, they become very difficult to spot and summarize in a pattern. Therefore, it is a satisfactory classification that 54 and 55 are able to be clustered as an individual class. However, the conclusion cannot be generalized that ART2 is able to always successfully spot a small volume of data in a large volume, since it needs to be tested with more data.

Classes V and VI are considered next. Class V consists of four samples, samples 48 (refinery R-JN) and 52 (refinery R-JM) as well as 56 and 59 (refinery R-DL). Class VI has a similar composition: samples 49 (refinery R-JN) and 53 (refinery R-JM) as well as 57 and 58 (refinery R-DL). We would rather regard them as a single class. It is well accepted that data clustering should only provide an estimation, and the final decision should be made by domain experts. We would like to point out that this result is consistent with the result obtained using the principal component analysis (PCA),[31] as shown in Figure 6. In Figure 6, the cluster represented by 4, 7, and 8 is the counterpart of class (V + VI). It needs to be pointed out that Figure 6 gives only 52 data samples.[31]

If we decrease the vigilance value, fewer classes will be formed, as shown in Tables 4 ($\rho = 0.9954$) and 5 ($\rho = 0.99504$). In both cases, class I always consists of 27 samples, representing the crude oils of Daqing. However, some other classes are mixed. For instance, class II in Table 3 represents only Xinjiang crude oil origin, while class II in Table 5 includes crude oil origins of Xinjiang (samples 28−40), Yumen (samples 41 and 42), Linshang (samples 48 and 49), Nanning (samples 52 and 53), and Renqiu (samples 56−59).

**Comparison with Other Methods.** Categorizing classification approaches into *supervised* and *unsupervised* is very useful for comparing vaious methods. Supervised classification requires sufficient training data sets with which there is a full understanding. After training, it can be used for prediction for unknown data. Since supervised learning is to learn from known data and predict for new data, there is always the danger that, when new data are not covered by the range of training data, the prediction may not be reliable. Back-propagation neural networks used by many researchers for infrared spectra interpretation belong to supervised classification.[2−4,13−17]

If there are not sufficient known data to be used for training, we have to use unsupervised techniques. The unsupervised approach is to discover from data which are different from supervised methods which learn from known to predict unknown.

Ordination methods such as principal component analysis (PCA)[34] and multidimensional scaling are unsupervised methods; however, they also have their problems.[35] With some sets of data, these methods may not give an adequate representation in two or three dimensions, and so visual

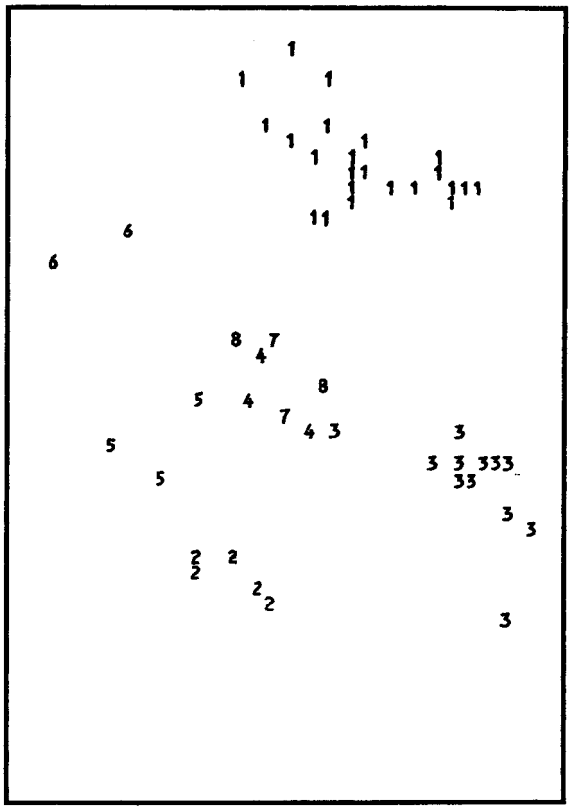

**Figure 6.** Classification using principal component analysis.

**Table 4.** Classification Results When the Vigilance Parameter $\rho = 0.9954$

| class | no. of samples | samples |
|---|---|---|
| I | 27 | 1−27 |
| II | 14 | 28−34, 36−42 |
| III | 7 | 35, 44−47, 50, 51 |
| IV | 1 | 43 |
| V | 8 | 48, 49, 52, 53, 56−59 |
| VI | 2 | 54, 55 |

**Table 5.** Classification Results When the Vigilance Parameter $\rho = 0.99504$

| class | no. of samples | samples |
|---|---|---|
| I | 27 | 1−27 |
| II | 22 | 28−34, 36−42, 48, 49, 52, 53, 56−59 |
| III | 7 | 35, 44−47, 50, 51 |
| IV | 1 | 43 |
| V | 2 | 54, 55 |

examination may not be possible. Sammon[36] gives an example where data generated to contain five groups in four dimensions are projected into the space of the two principal eigenvectors. Visual examination of this projection shows only four groups, since two of the clusters overlap completely in the two-dimensional space. The Kohonen network is similar to ART2 in the sense that both can continuously update the descriptions of classes during learning. However, the Kohonen network requires the number of classes to be fixed before learning starts. In contrast, the ART2 neural network determines the number of classes during learning as well as their descriptions.

## 4. CONCLUDING REMARKS

The 59 data samples were collected from 12 refineries having varied processing technologies and strategies, so samples from any two refineries have variations in properties. The ART2 neural network is able to ignore the slight variations and capture the main feature differences in all 59 samples to give seven classes. Analysis shows that five of the classes perfectly match the crude oil origins. An exceptional case (sample 35) is in agreement with prior observation of the spectrum, which further verifies the approach. There are two classes estimated by ART2 which cannot be satisfactorily explained. But when grouping them together, a reasonable cluster is formed, and it is consistent with the estimation by PCA.

The main advantage of ART2 neural network is that it is an *unsupervised* machine learning approach. Unlike BPNN, it does not need known and classified data sets as training data since it learns from unknowns, identifying the similarities of samples and predicting number of classes as well as data sample assignments.

## REFERENCES AND NOTES


(1) Clark, H. A.; Jurs, P. C. Classification of crude oil chromatograms by pattern recognition techniques. *Anal. Chem.* **1979**, *51*, 616−623.

(2) Klawun, C.; Wilkins, C. L. A novel algorithm for local minimum escape in back-propagation neural networks: application to the interpretation of matrix isolation infrared spectra. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 984−993.

(3) Klawun, C.; Wilkins, C. L. Joint neural network interpretation of infrared and mass spectra. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249−257.

(4) Munk, M. E.; Madison, M. S.; Robb, E. W. The neural-network as tool for multispectral interpretation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231−138.

(5) Kawahara, F. K. Identification and differentiation of heavy oil and asphalt pollutants in surface waters by comparative ratios of infrared absorbances. *Environ. Sci. Technol.* **1969**, *3*, 150−153.

(6) Mattson, J. S. "Fingerprinting" of oil by infrared spectrometry. *Anal. Chem.* **1971**, *43*, 1872−1873.

(7) Mattson, J. S.; Mark, H. B.; Kolpack, R. L.; Schutt, C. E. A rapid, nondestructive technique for infrared identification of crude oils by internal reflection spectrometry. *Anal. Chem.* **1970**, *42*, 234−238.

(8) Kowalski, B. R.; Bender, C. F. The K-nearest neighbour classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal. Chem.* **1972**, *44*, 1405−1411.

(9) Pichler, M. A.; Perone, S. P. Computer pattern recognition applications to chemical analysis. *Anal. Chem.* **1974**, *46*, 1790−1802.

(10) Brown, C. W.; Lynch, P. F. Infrared analysis of weathered petroleum using vacuum technologies. *Anal. Chem.* **1976,** *48*, 191−195.

(11) Mattson, J. S.; Mattson, C. S.; Spencer, M. J.; Starks, S. A. Multivariate statistical approach to the fingerprinting of oils by infrared spectrometry. *Anal. Chem.* **1977**, *49*, 297−302.

(12) Frankel, D. S. Pattern recognition of Fourier transform infrared spectra of organic compounds. *Anal. Chem.* **1984**, *56*, 1011−1014.

(13) Sutter, J. M.; Jurs, P. C. Neural network classification and quantification of organic vapours based on fluorescence data from a fibre optic sensor array. *Anal. Chem.* **1997**, *69*, 856−862.

(14) Wessel, M. D.; Jurs, P. C. Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks. *Anal. Chem.* **1994**, *66*, 2480−2487.

(15) Robb, E. W.; Munk, M. E. A neural network approach to infrared spectrum interpretation. *Mikrochim. Acta [Wien]* **1990**, *I*, 131−155.

(16) Munk, M. E.; Madison, M. S.; Robb, E. W. Neural networks models for infrared spectrum interpretation. *Mikrochim. Acta [Wien]* **1991**, *II*, 505−514.

(17) Fessenden, R. J.; Gyorgyi, L. Identifying functional groups in IR spectra using neural network. *J. Chem. Soc., Perkin Trans.* **1991**, *2*, 1755−1762.

(18) Zupan, J.; Novic, M.; Gasteiger, J. Neural networks with counterpropagation learning-strategy used for modelling. *Chemom. Intell. Lab. Syst.* **1995,** *27*, 175−187.

(19) Zupan, J.; Gasteiger, J. Neural networks for chemists: an introduction; Verlag Chemie: Weinheim, 1993.

(20) Novic, M.; Zupan, J. Investigation of infrared spectra-structure correlation using Kohonen and counterpropagation neural networks. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454−466.

(21) Knight, K. Connnectionist ideas and algorithms. *Commun. ACM* **1990**, *33*, 59−74.

(22) Schuur, J. H.; Selzer, P.; Gasterger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.

(23) Grossberg, S. Adaptive pattern classification and universal recording: I. parallel development and coding of neural feature detector. *Bio. Cybernet* **1976**, *23*, 121−134.

(24) Grossberg, S. Adaptive pattern classification and universal recording: II feedback, expectation, olfaction, illusions. *Cybernet* **1976**, *23*, 187−202.

(25) Carpenter, G. A.; Grossberg, S. ART2: self-organisation of stable category recognition codes for analogue input patterns. *Appl. Opt.* **1987**, *26*, 4919−4930.

(26) Carpenter, G. A.; Grossberg, S. A massive parallel architecture for a self-organising neural pattern recognition machine. *Comput. Vis. Graphics Image Process* **1987**, *37*, 54−115.

(27) Carpenter, G. A.; Grossberg, S. The ART of adaptive pattern recognition by a self-organising neural network. *Computer* **1988**, *21*, 77−88.

(28) Carpenter, G. A.; Grossberg, S. ART3: hierarchical search using chemical transmitters in self-organising pattern recognition architectures. *Neural Netw.* **1990**, *3*, 129−152.

(29) Caudil, M. Neural network primer, Part VIII. *AI Expert* **1989**.

(30) Wasserman, P. D. *Neural computing: theory and practise*; Van Nostrand Reinhold: New York, 1989.

(31) Zhou, H. Pattern recognition of lubricating base oils (in Chinese). M.Sc. Dissertation, Research Institute of Petroleum Processing, SINOPEC, 1988.

(32) Zhou, H.; Li, K.; Wang, X.; Xu, Y. F.; Shen, F. Pattern recognition studies on the influence of chemical composition on some properties of lubricating base oils. *Proceedings of the International Conference on Petroleum Refining Petrochemical Processing*, Beijing; 1991; Vol. 1, pp 387−392.

(33) Lynch, P. F.; Brown, C. W. Identifying source of petroleum by infrared spectroscopy. *Environ. Sci. Technol.* **1973**, *7*, 1123−1127.

(34) Yin, L. B. Classification results of mass spectra compounds by class modelling. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 1232−1234.

(35) Everitt, B. *Cluster analysis*, 2nd ed; Heinemann Educational Books Ltd.: Oxford, 1980.

(36) Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **1969**, *C18*, 401−409.


CI970434M