Journal of Chromatography, 126 (1976) 249–256
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

CHROM. 9378

ANALYSIS OF PYROLYSIS GAS CHROMATOGRAMS USING PATTERN RECOGNITION TECHNIQUES

E. KÜLLIK, M. KALJURAND and M. KOEL

Institute of Chemistry, Academy of Sciences of the Estonian SSR, Tallinn (U.S.S.R.)

SUMMARY

A set of 120 polymers was analyzed by using pattern recognition methods such as the cluster analysis, *K*-nearest neighbour and linear learning machine methods. A two-dimensional display of multivariate data is used to illustrate the results.

INTRODUCTION

The use of pyrolysis gas chromatography as a powerful analytical technique has been demonstrated many times[1–3]. However, its possibilities have not been exhausted and its wider application as an analytical method is complicated for several reasons. As an empirical method, the disadvantage of pyrolysis gas chromatography is that for the identification of unknown substances a compilation of pyrograms must be available. Depending on the conditions of pyrolysis and gas chromatographic analysis, different results may be obtained by different laboratories.

A possible means of analyzing empirical results is to apply the pattern recognition technique, which seems to be very useful for handling pyrolysis gas chromatographic data. Firstly, the pattern recognition method enables one to obtain information about a substance of interest when no reference chromatograms are available, and secondly, it can tolerate the existence of deviations in the initial data. It also enables one to overcome the problem of deviations in the retention times of corresponding pyrolysis products from different polymers. The second characteristic provides the possibility of reconciling differences in results obtained by different laboratories.

In a previous study[4], we described one of the pattern recognition methods (the linear learning machine method) used to identify different molecular groups in fibres. The task in the present work was to continue studies on the application of this method. The theory of pattern recognition is not given here because sufficient information is available elsewhere[5,6].

EXPERIMENTAL

Unfortunately, the pyrolysis gas chromatograms presented in many scientific papers were obtained under different conditions, and it was difficult to find suitable chromatograms. We therefore used chromatograms produced in our laboratory. The set of 120 polymers selected consists mainly of several natural and synthetic fibres. All fibres were pyrolyzed by using a Curie-point Pye Unicam pyrolyzer at 980°. A Perkin-Elmer Model 900 gas chromatograph was used for the analysis of the thermal degradation products.

The other operating conditions were: dual flame-ionization detector; Inerton AW-DMCS/Carbowax 20M column, length 2 m, I.D. 2 mm; initial column temperature 70°, programmed at 6.5°/min to a final temperature of 190°, which was maintained for 10 min.

For data processing, an off-line computer system was used. The gas chromatographic output signal was handled by using a Hewlett-Packard 3370A integrator and a Videoton 1010B computer. The retention time and peak area of every peak were punched on to paper tape.

PRESENTATION OF PYROGRAMS TO THE COMPUTER

The most precise method for the presentation of pyrolysis gas chromatograms to the computer is to give the name (retention time) and intensity (peak area) of each individual thermal degradation product for all pyrograms available. The best means of solving this problem is to identify all individual components in the pyrogram. For the calculation, each pyrogram is presented as a sequence of numbers, every peak being characterized by numbers that are in accordance with the peak intensity. If there is no peak, the intensity is zero. Such a presentation is commonly used for plotting gas chromatograms in tabular form. Unfortunately such a method of coding is limited by the set of substances used and needs a large computer memory.

In our study, peaks were coded as in infrared or mass spectroscopy. Each pyrogram was divided into zones of equal width and the most intense peak was taken into account. All chromatograms were presented to the computer as a normalized and logarithmic sequence of intensities of selected peaks. For the calculation, the integer part of the logarithmic value was taken.

Pyrolysis gas chromatograms were divided correspondingly into 20 and 40 zones, giving two sets of data (Fig. 1). Such "low" and "high" data sets permit one to estimate the loss of information by coding.

An essential parameter is the ratio between the number of objects presented and the number of measurements that characterize a particular object. Let us denote this ratio by s. From the literature[6] it is known that $s \geqslant 3$ is needed. In our work, the conditions $s_{20} = 6$, $s_{40} = 3$ were followed. According to the general geometric concept, coded programs in pattern recognition are treated as a set of points in either 20- or 40-dimensional hyperspace.

CLUSTER ANALYSES OF PYROGRAMS

It is interesting to elucidate groups of similar pyrolysis gas chromatograms in the given set of data, *i.e.*, to make a classification of polymers on the basis of their pyrolysis chromatograms. In a geometric sense, one looks for a set of points (clusters) in the hyperspace. We used the Euclidean distance between two points for as a measure of similarity; for finding clusters, two algorithms were used.

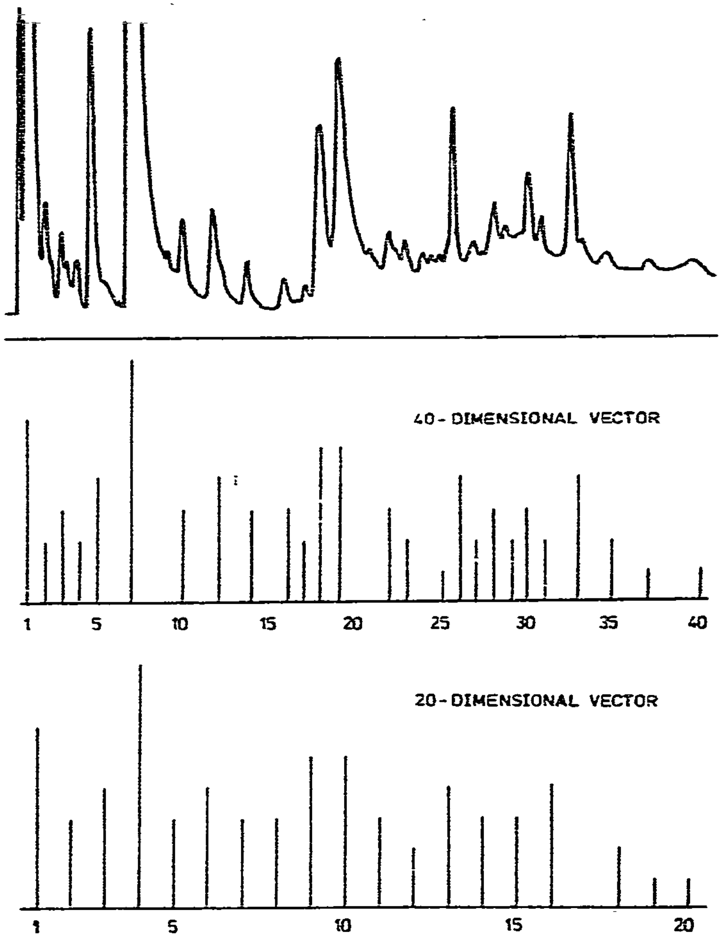

Fig. 1. Coding a pyrogram in the 40- and 20-dimensional cases.

In the first algorithm[7], the number of clusters is not fixed and a parameter R is assigned to determine the number of clusters; the points that are in one hypersphere, of radius R, form a cluster. The centre of the hypersphere is located at the maximum density of the points. If the number of clusters is constant for different R values, then there are well separated and compact clusters in the hyperspace.

Another algorithm may start from an arbitrary point. All points that are nearer to that point than the value of the threshold, r, are included in a cluster. According to the measure of similarity, r new points are included with these points in a cluster. The procedure is repeated until there are no points to add to this cluster, then a new point that does not belong to the cluster so formed is taken and the above procedure repeated. In this way, one can connect all of the existing objects to the clusters.

These two techniques gave virtually identical results for both the 40- and 20-dimensional cases. Three large clusters are formed; one consists of rubbers and the second of polyacrylonitrile fibres. The third cluster consists of $-CH_2-CHR$-type fibres, where R is an alcoholic hydroxyl acetate radical or chlorine group. The other fibres (polyamides, polyurethanes, polyesters) have no tendency to form clusters, *i.e.*, they are homogeneously distributed in the hyperspace.

If the radius *R* or the similarity measure *r* was increased, then all three clusters converged into one large cluster, but the separately located points remained beyond the cluster as before. The parameters *R* and *r* must, of course, be changed within the appropriate limits, otherwise one obtains only one cluster or as many clusters as there are objects.

The results of the cluster analysis demonstrate that there is a cluster with a large number of members (about half of the inspected polymers), which can be divided into three smaller clusters, and a number of separately located substances.

CORRELATION STUDIES BETWEEN PYROLYSIS GAS CHROMATOGRAMS AND THE STRUCTURES OF SUBSTANCES

It is convenient to use pattern recognition techniques to establish whether or not a certain property of the object appears in the pyrolysis gas chromatogram. For classification purposes we used two methods, the *K*-nearest neighbour[8] and the linear learning machine methods[9]. If the *K*-nearest neighbour method is used, the substance is classified according to its *K*-nearest neighbours; in this work $K = 1$ and 3. Using the linear learning machine method, one calculates the dot product between the object vector and an appropriately derived weight vector, the object being classified according to the sign of the dot product. Geometrically, this means that the object is classified according to the side of the classifying hyperplane on which it lies.

The results are presented in Table I. We have a binary classification for each property, *i.e.*, the object may or may not have a particular property. Table I gives the percentage of the prediction that characterizes the probability of classification

TABLE I

PERCENTAGES OF PREDICTION

Property	*Positive class number*	*K-nearest neighbour method*				*Linear learning machine method*	
		K = 1		*K = 3*			
		20-dimensional	*40-dimensional*	*20-dimensional*	*40-dimensional*	*20-dimensional*	*40-dimensional*
Nitrogen	57	73	77	72	72	55	59
–CN	25	92	91	93	90	74	74
NH–CO	31	76	81	74	79	67	51
Benzene ring	20	86	83	90	83	61	75
Oxygen	61	73	83	72	75	58	55
O, N and benzene ring in main chain	50	72	82	67	77	59	48
–C=O	49	68	75	66	70	43	51
–O–C=O	26	77	74	77	77	81	74
–Cl	19	85	89	84	85	59	69
Polyolefin	14	92	87	85	84	84	84

correctness[10]. The column "positive class number" shows the number of polymers that have a particular property. The greater the prediction, the clearer the structural element of a polymer is expressed in the pyrogram.

Fukunaga and Olsen[11] described an interesting technique for the two-dimensional display of multivariate data, namely the so-called *d*-displays, and we used it to illustrate our results. The method displays the points on a two-dimensional display, the coordinates of which are the squared Euclidean distances from two particular points in the *n*-dimensional space. These two points are the geometric centres of two classes. The method preserves some geometric structure while placing a heavy weight on class separability. A straight line on the *d*-display corresponds to some decision boundary in the *n*-dimensional object space that is reduced to a hyperplane for the line with an angle of 45° with respect to the d_1 axis. Fig. 2 shows the *d*-display for five classes. The overlapping of the $-C{\equiv}N$, $=N-$ and $-Cl$ classes is considerably less than for the "hetero-atom in the main chain" class. Partial overlapping of two classes in the display does not mean that the two classes are not separated at all but that other method must be used to find the decision boundary, such as *e.g.*, the linear learning machine method or the data must be normalized appropriately[11].

DISCUSSION

Bearing in mind the specific features of pyrolysis gas chromatography, the results of the cluster analyses of the pyrolysis products of the polymers can be summarized as follows. Polyacrylonitrile, polyvinyl and rubber polymers give characteristic pyrograms that enable one to differentiate them from the others. However, it is difficult to differentiate the above polymers inside a given class. On the other hand, individual polyamide, polyether and polyurethane pyrograms are very characteristic and are easily differentiated from each other.

It is evident that our results depend on the experimental conditions used. Using another liquid phase with a polarity different from that of Carbowax 20M, one can obtain different results. As a practical consequence, we have a method for estimating the suitability of a particular liquid phase for the analysis of a particular class of polymers. In this work it seems to be valid for polyamide, polyester and polyurethane polymers. However, if the constituents of the polymers are very similar, the polarity change might have no effect. All this is also valid for the pyrolysis conditions.

The results in Table I show that there are several functional groups the existence of which can be established with reasonable probability (–CN, –Cl and benzene ring). There are some functional groups, however, for which the prediction is on the random guessing level (–CO, –O–). In the linear learning machine method, the random guessing level is 50%. If the use of the pattern recognition technique fails for some property, the two main reasons are that either the presentation (or method used) does not express the expected property, or the presentation is coarse.

As can be seen from Table I, the prediction for $K = 1$ (40-dimensional) is better than for $K = 1$ (20-dimensional) and $K = 3$ (40- and 20-dimensional). $K = 1$ (20-dimensional) and $K = 3$ (40- and 20-dimensional) give virtually identical results. Therefore, we can conclude that improving the resolution (as well as K) does not have a significant effect on the improvement of prediction. The percentage of prediction obtained characterizes the ability of the pyrograms to express the tested property. It can be seen from Fig. 2 that most of the classes overlap considerably and on increasing the dimension of the hyperspace the overlapping of the classes decreases. It seems that they did not form the linearly separated set of points (as in our previous work[4]), which is why the results of the *K*-nearest neighbour method are better than those of the linear learning machine method. On the other hand, it is evident that the results obtained with pattern recognition techniques depend on the pre-processing method, *i.e.*, the way in which the pyrograms are presented to the computer. In this work, we did not attempt to obtain a good separation for a particular class. There are effective pre-processing and feature selection methods in pattern recognition for improving inter-class resolution[5] and one can obtain considerably better results than those in Table I for a particular problem of interest. Our presentation (logarithmic intensities), which places a heavy weight on the smaller peaks, may emphasize unimportant features for a particular class. It is clear from the above discussion that pre-processing methods for pyrograms need further investigation.

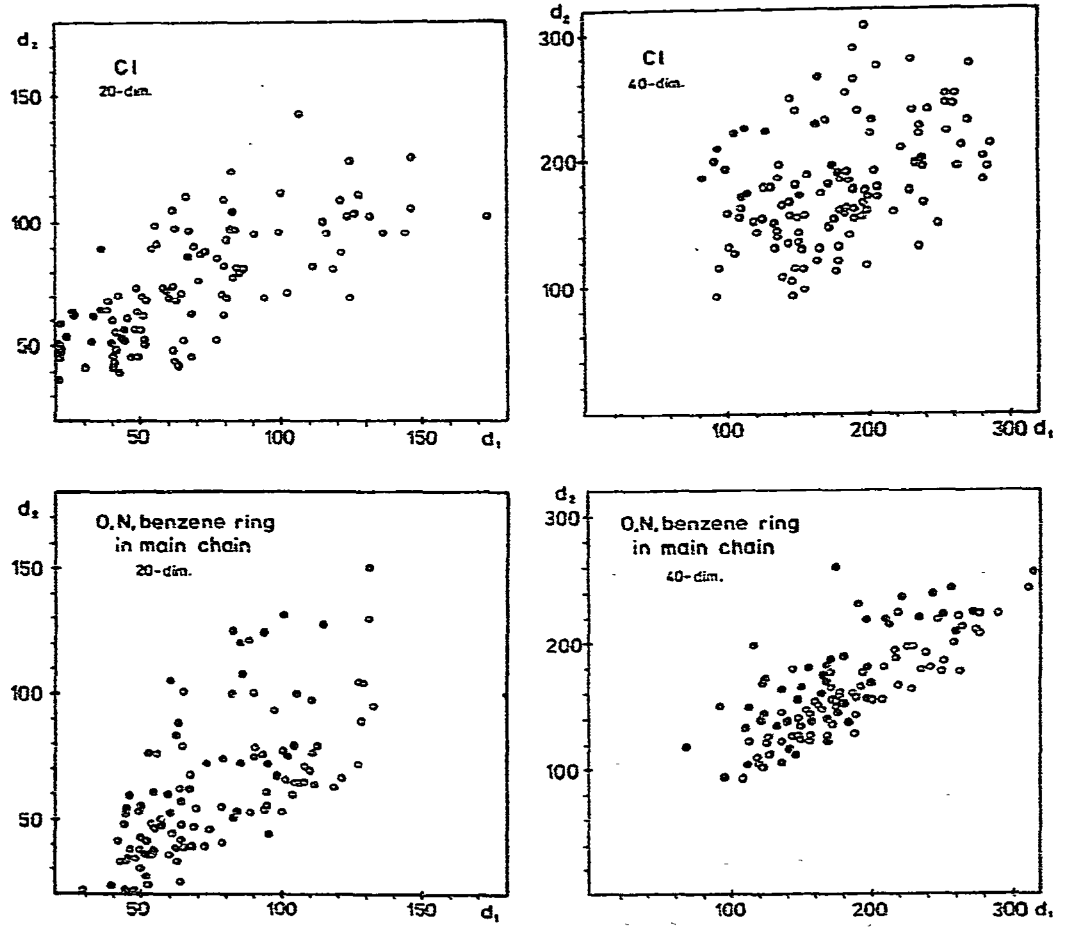

Fig. 2.

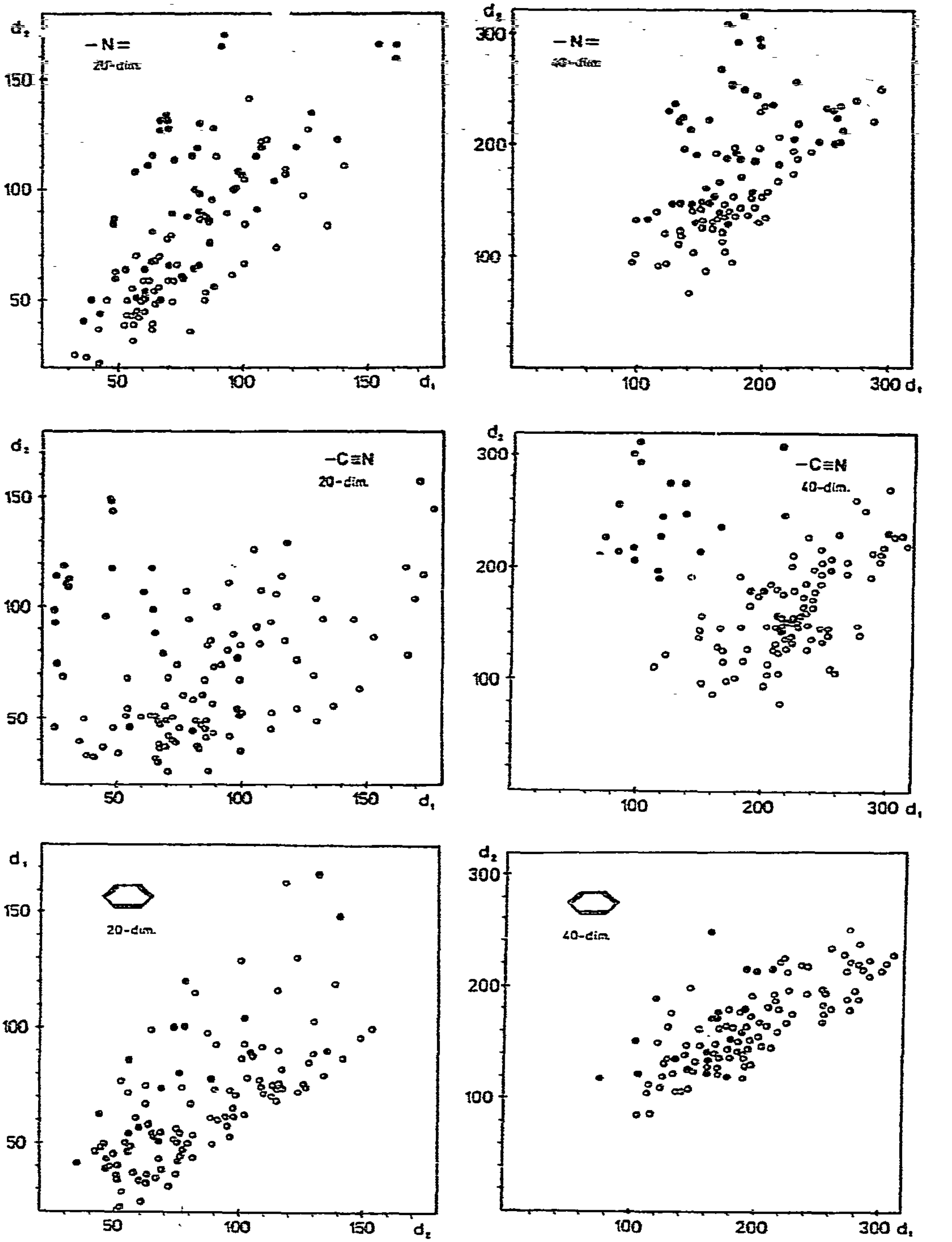

Fig. 2. *d*-Displays for some tested properties. d_1- and d_2-squared distances from the geometric centres of the positive and negative classes. ●, Positive class members; ○, negative class members.

The interpretation of spectral data with pattern recognition is a completely empirical method, and therefore only practical results will show which chemical structures are classified well.

REFERENCES

1 P. G. Simmonds, G. P. Shulman and C. H. Stembridge, *J. Chromatogr. Sci.*, 7 (1969) 36.
2 J. Janák, *Collect. Czech. Chem. Commun.*, 25 (1960) 1780.
3 L. S. Ettre, *J. Chromatogr.*, 112 (1975) 1.
4 E. Küllik, M. Kaljurand and M. Koel, *J. Chromatogr.*, 112 (1975) 297.
5 J. M. Mendel and K. S. Fu (Editors), *Adaptive, Learning and Pattern Recognition Systems*, Academic Press, New York, 1970.
6 B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 94 (1972) 5632.
7 V. N. Elkina and N. G. Zagoruiko, *Vychislitel. Sist.*, 22 (1966).
8 T. M. Cover and P. E. Hart, *IEEE Trans. Inf. Theory*, IT-13 (1967) 21.
9 N. J. Nilsson, *Learning Machines*, McGraw-Hill, New York, 1965.
10 T. L. Isenbour and P. C. Jurs, *Anal. Chem.*, 43 (1971) 20A.
11 K. Fukunaga and D. R. Olsen, *IEEE Trans. Comput.*, C-20 (1971) 917.