ELSEVIER

Polymer Testing 21 (2002) 43–47

POLYMER TESTING

www.elsevier.com/locate/polytest

Property Modelling

Application of cluster analysis method in prediction of polymer properties

Barbara Dębska [a,*], Elwira Wianowska [b]

[a] *Department of Computer Chemistry, Rzeszów University of Technology, 6 Powstańców Warszawy Avenue,*
35-041 Rzeszów, Poland
[b] *Department of Science of Materials and Shoe Technology, Radom University of Technology, 22 Malczewskiego,*
26-600 Radom, Poland

Received 22 February 2001; accepted 17 April 2001

Abstract

This paper presents the results of the experiments whose purpose was to examine a new polymeric material based on acrylamide-modified melamine resins. Acrylamide was used to enhance the waterproof properties of each resin examined. To prepare the process of production for these resins, a series of experiments on the laboratory scale was carried out to provide sufficient information concerning the characteristics of the proposed products. A large quantity of data was dealt with, a computer being the tool for building a database of the results of the chemical processes being carried out. Interpretation of the stored data enabled us to discover regularities existing within the data and to draw conclusions from this information. The cluster analysis method was used for determination of the origin of the samples (objects), and for prediction of properties of acrylamide-modified melamine resins. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Application of system science in chemistry; Data clustering; Data classification

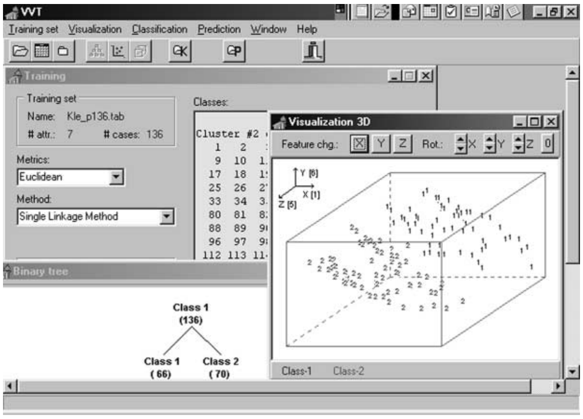
1. Introduction

Cluster analysis provides a number of methods to obtain an insight into data sets, and to extract relevant information from them. This analysis is the most important tool used for interpreting multivariate data containing objects and features, and sometimes also properties. Cluster analysis generically refers to different multivariate methods designed to create homogeneous sets of objects called cluster. An object is any real item such as a sample, a chemical structure, or a technological process. An object is characterised by a set of features. A feature is a numerical variable (in this paper — chemical-analytical measured data) such as concentration, temperature, or reaction time. These data are best described by an $(n{\times}p)$ matrix X, containing a row for each of the n objects, and a column for each the p features and/or properties [1]. Each object then corresponds to a point in the p-dimensional feature space.

Graphic presentation of the data structure of d-dimensional picture area as two- or three-dimensional drawings may be helpful in the evaluation of the obtained results. For visualisation of multivariate objects special numerical algorithms are widely used. Visual inspection of a figure (see Fig. 1) often indicates clusters of similar objects as well as other relationships within the data. The relationships between clusters are shown in form of a dendrogram or a table. The distance between analysed objects is considered the measure of their similarity. The distance and similarity are therefore reciprocal. The distance between objects in a p-dimensional space can be defined among others as the Euclidean distance, City Block (called also Manhattan) distance, the Tshebysh-

* Corresponding author. Tel.: +48-17-86-51-358; fax: +48-17-85-41-519.
E-mail address: bjdebska@prz.rzeszow.pl (B. Dębska).

PII: S0142-9418(01)00045-9



Name: Kle_p136.tab

attr.: 7 # cases: 136

Metrics: Euclidean

Method: Single Linkage Method

```
Cluster #2
   1   2
   9  10  1.
  17  18  1
  25  26  2
  33  34  3.
  80  81  8
  88  89  9
  96  97  9
 112 113 11
```

Fig. 1. The results of the cluster analysis (table, part of dendrogram, and 3-d picture).

ev's distance, the Minkowski distance, the Hamming distance, or Tanimoto distance [2]. In the presented research, cluster analysis is based on calculating distances with the Euclidean, City Block and Tshebyshev's methods.

The usual aims of data interpretation are: search for clusters containing similar objects and investigation of relationships between the set of features and the set of properties. A typical cluster analysis consists of the following steps: (i) collecting data for known objects, that is for the so-called training set, (ii) classification of the objects constituting the training set, and (iii) applying the results of step ii to the classify new objects and to predict their properties.

The research on cluster-creation methods involved analysing various algorithms of grouping objects in non-empty, separate classes. Marek [3] presents a general division of cluster analysis methods. Among the methods described in the aforementioned paper, the most often used ones are the hierarchical methods, which separate objects into hierarchical clusters, i.e. where some clusters contain other clusters, from many clusters containing one object to one cluster containing all the objects. Hierarchical clustering methods involve building a hierarchical classification of the objects in the dataset by a series of either binary divisions or agglomerations. These methods can realise either a descending algorithm (where a set of objects is divided into two subsets, then each subset is divided into two successive ones, and so on) or an ascending (agglomerative) algorithm (where the starting point is a single object treated as a one-element cluster; objects are grouped in successive clusters generating a cluster tree — a dendrogram).

In the case of an ascending algorithm, the more similar the values of objects' features are, the sooner the objects are grouped. This procedure has been used for a long time for empirical research in natural science (e.g. in data mining). For that reason, we've undertaken research on the efficiency of hierarchical algorithms of object grouping. This paper discusses some chosen methods of cluster analysis (the so called SAHN group, that is Sequential, Agglomerative, Hierarchical and Non-overlapping methods) which are particularly useful for object classification in chemical science and other, related fields of knowledge.

2. SAHN methods

A number of good commercial software products for classification and analysis of multivariate objects (observations) (e.g. the program Cluster Analysis in the STATISTICA system [4]) are available today. These products apply various methods of cluster analysis. For our (non-commercial) computer system SCANKEE (in the VVT module) [5] we've chosen the SAHN methods, namely: (1) SLM — single linkage method, (2) GAM — group average method, (3) WAM — weighted average method, (4) CLM — complete linkage method, (5) UCM — unweighted centroid method, (6) WCM — weighted centroid method (median method), (7) minimum variance method (Ward method), and (8) flexible SAHN strategy [3]. The differences of the various methods (algorithms) lie in how they group together (link) objects, or progressively build clusters based on similarities between the objects. For example, in the algorithm:

- SLM [6], the consecutive agglomerations are based on measuring the distance to the nearest object in a group,
- CLM [7], the linkage rule is still based on smallest distance, but the distances between clusters are calculated on the basis of the furthest objects, or
- Ward's method [8,9], it is a method in which the variance of groups is assessed during the clustering process, and the group which experiences the smallest increase in variance with the iterative inclusion of an object will receive that object.

The closer to each other the objects are situated, the bigger is their similarity. In our VVT module for cluster analysis, we use three metrics to calculate distances: Euclidean, City Block, and Tshebyshev's. The most popular distance measure in multidimensional space is based on the Euclidean formula:

$$d_{\mathrm{Euclidean}}=\left[\sum_{i=1}^{p}(x_i-y_i)^2\right]^{1/2}$$

The City Block distance is used relatively often, as well:

$$d_{\mathrm{City\ Block}}=\sum_{i=1}^{p}|x_i-y_i|$$

Sometimes, the one in use is the Tshebyshev's distance:

$$d_{\text{Tshebyshev}} = \max_{i=1,\ldots,p} |x_i - y_i|$$

3. Preparation of polymers

Hydroxymethyl derivatives of acetone react easily with melamine. The reaction of polycondensation proceeds in a resin-like medium, in an elevated temperature and in the presence of acidic or basic catalysts. It is conducted with the use of reactive hydroxymethyl groups and amine groups (deriving respectively from acetone and melamine). The products of a crosslinking reaction of melamine resins are applied as adhesives for laminates and as binders for chipboards, failing to provide sufficient waterproof properties to the materials concerned. These properties can be improved by raising the temperature of the process of production and extending the time of pressing, which augments, however, release of free formaldehyde. To improve the waterproof properties of the resins examined, it was therefore decided to copolymerize melamine and formaldehyde with acrylamide [10,11]. Addition of acrylamide as a reactive reagent enables us to create a product of a higher degree of crosslinking and of higher use value, which is possible because acrylamide, having double bonds in its structure, is able to react according to the radical polymerisation mechanism. The rate of the polymerisation process, and consequently the rate of processing, depend in this case entirely on the rate of decomposition of the radical initiator H_2O_2.

Acrylamide-modified melamine resins are obtained in a two-stage process. The product of the first stage is a reactive melamine solvent, obtained in a reaction of acetone with 10-fold excess of formaldehyde. The second stage consists in: (a) dissolving melamine in the reactive melamine solvent (i.e. in hydroxymethyl derivative of acetone) and (b) adding (consecutively) acrylamide as the comonomer.

One way of examining waterproof properties of a given resin is to check its loss of weight in the process of curing. The loss of weight of the acrylamide-modified melamine resin was examined after samples of the resin had been heated for 30 min in boiling water, the curing time being 373 K for half of the samples and 393 K for the other half. Within 20 to 50 min, protraction of the curing time was found to result in advantageously smaller weight loss. The weight loss was found to be inversely related to the curing time. The loss of weight is calculated as follows:

$$Ub = (a - c)/a \times 100\%$$

where a=sample's mass before determination, g; c=sample's mass after being boiled and cured, g.

The results of the experiments are coded based in the form of a specific formalism of knowledge about technological processes (i.e. knowledge association formalism):

physical and/or chemical polymer mixture + **technological parameters** → ← **the properties of the obtained product**

This formalism can be used to describe other technological processes. It is used, for example, in computer-assisted engineering of materials and it enables us to solve two main problems that occur at the stage of planning a chemical experiment:

1. the problem of classification — whether a given polymer mixture, together with chosen technological parameters, will enable us to produce a good product, and
2. the problem of prediction — what polymer mixture and technological processes need to be applied to obtain a particular product.

The VVT module of the SCANKEE system, applied on this stage of our research, was used to solve the two problems.

4. Classification and prediction of polymers properties

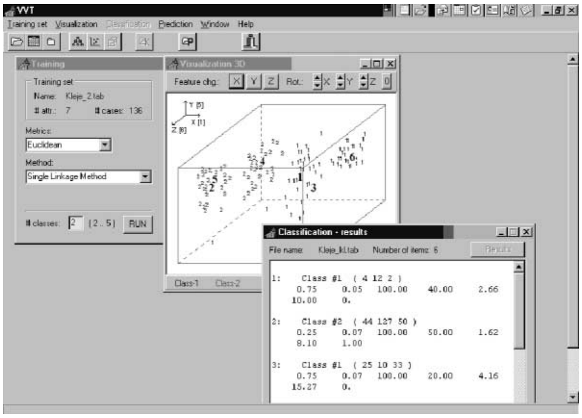
The VVT module has examined many other real chemical experimental data sets. Most of the data stored in the training set discussed in this paper derive from a chemical laboratory, and they are results of experiments on melamine resins. Each resin sample of this set is described by means of six numerical attributes: the amount of acrylamide, the amount of cross-linking agent, the temperature of the reaction, its duration, the amount of formaldehyde in the sample, and the resin mass decrement. This standard data set (a collection of analytical data for 136 resin samples) has the form of an ASCII file. It is a matrix of 136 rows and 6 columns, with an extra binary column of classifying features (CF, decision features defining the quality of the resins). A fragment of this base is presented in Table 1. The VVT module was used for graphical presentation of experimental data sets (see Fig. 1), for classification of resin samples (see Fig. 2), and for prediction of new resins, which would have some particular, desired properties (see Fig. 3).

Figure 1 presents the results of a cluster analysis. One can clearly notice two clusters of resin samples, divided according to their good or bad waterproof properties. The dendrogram presents the number of samples in each cluster (74/62).

An unknown object (a resin sample) can be classified by investigating k nearest-neighbouring (KNN) objects whose class memberships are known. In order to find the

Table 1
Fragment of laboratory database

AA (g)	H_2O_2 (%w)	Temp. (°C)	Time (min)	F (%)	Ub (%)	CF
0.75	0.05	100	20	5.35	15.30	0
0.75	0.05	100	30	2.68	13.64	0
0.75	0.05	100	40	3.56	13.82	0
0.75	0.05	100	50	2.60	12.51	0
0.75	0.05	120	20	1.63	10.21	0
0.75	0.05	120	30	1.85	9.37	1
0.75	0.05	120	40	1.19	7.20	1
0.75	0.05	120	50	0.98	6.50	1



Training set
Name: Kleje_2.tab
# attr.: 7 # cases: 136
Metrics: Euclidean
Method: Single Linkage Method
# classes: 2 [2..5]

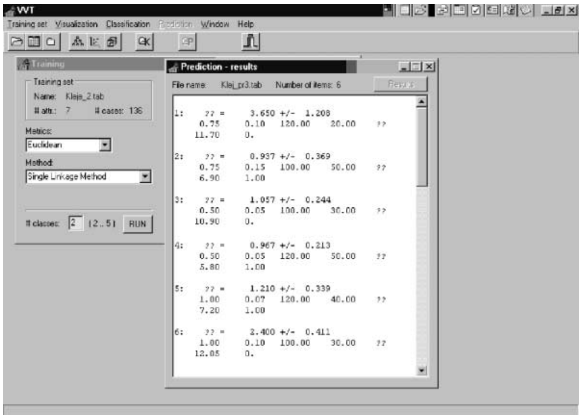
Classification - results
File name: Kleje_kl.tab Number of items: 6

```
1:    Class #1  ( 4 12 2 )
    0.75      0.05   100.00     40.00      2.66
   10.00      0.

2:    Class #2  ( 44 127 50 )
    0.25      0.07   100.00     50.00      1.62
    8.10      1.00

3:    Class #1  ( 25 10 33 )
    0.75      0.07   100.00     20.00      4.16
   15.27      0.
```

Fig. 2. Results of classification process.



Training set
Name: Kleje_2.tab
# attr.: 7 # cases: 136
Metrics: Euclidean
Method: Single Linkage Method
# classes: 2 [2..5]

Prediction - results
File name: Klej_pr3.tab Number of items: 6

```
1:    ?? =     3.650 +/-  1.208
    0.75      0.10   120.00     20.00     ??
   11.70      0.

2:    ?? =     0.937 +/-  0.369
    0.75      0.15   100.00     50.00     ??
    6.90      1.00

3:    ?? =     1.057 +/-  0.244
    0.50      0.05   100.00     30.00     ??
   10.90      0.

4:    ?? =     0.967 +/-  0.213
    0.50      0.05   120.00     50.00     ??
    5.80      1.00

5:    ?? =     1.210 +/-  0.339
    1.00      0.07   120.00     40.00     ??
    7.20      1.00

6:    ?? =     2.400 +/-  0.411
    1.00      0.10   100.00     30.00     ??
   12.05      0.
```

Fig. 3. Result of prediction of resins' properties.

nearest neighbours of the examined object, it is necessary to compute the distances between this object and all objects of the data set. The unknown object is assigned to the cluster that groups the majority of the *k* neighbours. To ensure explicit classification, *k* needs to be an odd number.

As Fig. 2 shows, the results of the classification process can be displayed on one screen both as a chart (here: results of classification for six new resins) and as a numerical scrolling list (here: first line of each description of a sample indicates the cluster it was classified to and three nearest neighbouring elements).

The prediction algorithm predicts a given searched property for a simulated object. This property is calculated on the basis of the values (weighted average) of the same property for 3-KNN objects, as follows:

$$p=p_1\times w_1+p_2\times w_2+p_3\times w_3$$

where: $w_i=d_i/(d_1+d_2+d_3)$, i=1, 2, 3; p_1, p_2, p_3=searched property's values for 3-KNN objects; w_1, w_2, w_3=weight coefficients for 3-KNN objects; d_1, d_2, d_3=distances between each of the 3-KNN objects and the simulated object.

Figure 3 displays the results of the prediction process for six simulated resins. The double question mark sign signifies the property being predicted, the first line of each description of a sample indicates its predicted value.

5. Conclusion

Much attention was paid to selecting the most effective SAHN methods (Table 2). In a preliminary numerical experiment that was carried out on the training set, Euclidean distance proved to be the most effective metric for cluster analysis. It was observed that some of the SAHN-procedures could also be successfully applied for classifying resin samples. The samples are separated into two classes, distribution error amounts to <10%. The distribution error was defined as a fraction of the total objects not correctly classified in their clusters.

Cluster analysis, object classification and prediction of properties for simulated objects are very useful for solving many chemical problems, e.g. in classification of materials, quality control, investigation of relationships between chemical compounds (composition of

Table 2
Results of testing the analytical data

Metric	Good results of cluster analysis (distribution error $<10\%$)	Bad results of cluster analysis (distribution error $\geq 10\%$)
Euclidean	single linkage method, group average method, weighted average method, unweighted centroid method	complete linkage method, weighted centroid method, minimum variance method, flexible SAHN strategy
City Block	–	all methods
Tshebyshev's	–	all methods

substrates) and their properties, or optimisation of chemical technological processes.

References

[1] K. Varmuza, Chemometrics: multivariate view on chemical problems, in: P.R. Schelyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer III, P.R. Schreiner (Eds.), The Encyclopaedia of Computational Chemistry, vol. 1, John Wiley and Sons, Chichester, 1998, pp. 346–366.

[2] Z.S. Hippe, Finding nearest neighbours for symbolic attributes: A comparison of selected methods, in: Proceedings of the CAI'98 Conference, Lódz, 1998, pp. 28–30.09.

[3] T. Marek, Cluster Analysis in Empirical Researches, PWN, Warsaw, 1989.

[4] STATISTICA for Windows (vol. III), Statistics II, StatSoft Inc., 2300 East 14 Street, Tulsa, OK 74104, USA, 1999.

[5] Z.S. Hippe, B. Dębska, M. Mazur, New programming tool for real-world applications of CAE in industry, in: K.E. Oczo, R. Ostholt (Eds.), CAE Techniques PRz, Rzeszów, 1994, pp. 321–327.

[6] P. Sneath, The application of computers to taxonomy, J. Gen. Microbiol. 17 (1957) 201–226.

[7] R. Sokal, C.D. Michener, Statistical method for evaluating systematic relationship, Univ. Kan. Sci. Bull. 38 (1958) 1409–1438.

[8] J.H. Ward, Hierarchical grouping to optimise an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.

[9] J.H. Ward, M.E. Hook, Application of a hierarchical grouping procedure to a problem of grouping profiles, Educ. Psychol. Meas. 23 (1963) 69–82.

[10] B. Dębska, E. Malinowska, Studies on the properties of acrylamide-modified melamine resins: Preparation of polymers and statistical evaluation of experimental data, Polimery, Warsaw 44 (9) (1999) 608–613.

[11] Dębska B., Wianowska E., Acrylamide as agent modifying melamine–acetone–formaldehyde resins, Polymer Testing 21 (2002) 49–55.