ESTIMATION OF THE STABILITY OF COMPLEX HYDRIDES BY A PATTERN-RECOGNITION CLASSIFICATION METHOD

Jiří FUSEK and Oldřich ŠTROUF

Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, 250 68 Řež

Received July 21st, 1978

The ABH_nD_{4-n} complex hydrides (A = alkali metal, B = the IIIB Group atom, n = number of hydride atoms and D = ligand) were classified as stable or unstable ones by a pattern—recognition method. The method is based on a measure of similarity which is represented by the ratio of squared distance of the object from the centre of the class and the averaged squared distance of this centre from all prototypes in the class. The distances are measured in transformed and normalized spaces for each class, the individual class approaching a hypercube with an edge length (standard deviation) of one. The results are directly compared with those previously obtained by SIMCA principal components method using the same data base (MODEL-1).

Pattern recognition is a rapidly developing methodology of data analysis which has been recently used for classification purposes[1-5]. It consists, principally, of two main parts: feature selection and classification. Both parts represent very extensive fields. Moreover, these principal parts are frequently accompanied by additional ones (*e.g.* the preprocessing of data and the display of the output) which could improve and/or facilitate the pattern recognition analysis under study. Hencefore, it is not surprising that the definition of pattern recognition is not sufficiently rigid (see *e.g.* the discussion in ref.[6]). In such a situation the choice of the most convenient method for a practical problem is rather difficult. We believe that the direct comparison of pattern recognition methods on the basis of identical data base remains an important way of the testing in spite of the recent appearance of comparative theoretical means *e.g.* those of the information theory[7]. Therefore, the direct comparison is used also in this paper.

Recently, pattern-recognition methods have been extensively applied in the interpretation of multivariate chemical data[8-10]. In this paper we compare directly the results of two different classification methods using identical data for complex hydrides of the general formula ABH_nD_{4-n} as compiled in the "Data Base I" (ref.[11]) under the name MODEL-1. The first of the methods is the Wold's SIMCA method[12,13], the results of which have been described recently[14], the second new one, is described here in detail.

The SIMCA method is based on the description of similar objects in an individual class q by means of data $y_{ik}^{(q)}$ using the analogy model of the principal components form

$$y_{ik}^{(q)} = \alpha_i^{(q)} + \sum_{a=1}^{A} \beta_{ia}^{(q)} \theta_{ak}^{(q)} + \varepsilon_{ik}^{(q)} . \tag{1}$$

The parameters $\alpha_i^{(q)}$, $\beta_{ia}^{(q)}$ and $\theta_{ak}^{(q)}$ are determined on the basis of the experimental data matrix, so as the sum of squared residuals $\varepsilon_{ik}^{(q)}$ be minimal.

Our classification method can be described by the following algorithm:

1) Start with the preprocessing step in which the data of the training sets are autoscaled so that the mean be zero and standard deviation be unity according to

$$y'_{ik} = \frac{y_{ik} - \bar{y}_i}{(\overline{y_i^2} - \bar{y}_i^2)^{1/2}}\,, \qquad (2)$$

where y'_{ik} are autoscaled data for prototypes k, y_{ik} are the original data, $\bar{y}_i$ is the mean of the data for a variable (dimension) i, and $\overline{y_i^2}$ is the mean of squared data.

2) The covariance matrix **U** is formed for the individual training set with the elements

$$u_{ir} = u_{ri} = \frac{\sum_{j=1}^{N_q} y'_{ij} \cdot y'_{rj}}{N_q} - \frac{\sum_{j=1}^{N_q} y'_{ij} \cdot \sum_{j=1}^{N_q} y'_{rj}}{N_q^2}\,, \qquad (3)$$

where $i, r = 1, 2, \ldots,$ R and R is the number of variables. N_q is the number of the prototypes in the q-th class.

3) For each class, the eigenvectors $\mathbf{e}_i$ are calculated from the matrix **U**

$$\mathbf{e}_i \,.\, \mathbf{U} = \mathbf{e}_i \,.\, \lambda_i \qquad (4)$$

4) The autoscaled data of the prototypes in the class are transformed by means of the matrix of $\mathbf{e}_i$ according to

$$x_{ij} = \sum_{r=1}^{R} y'_{rj} \cdot \mathbf{e}_{ir} \qquad (5)$$

5) The eigenvalues λ_i calculated by Eq. (*4*) represent variances s^2 of the corresponding dimensions after transformation. We consider all $\lambda_i^{1/2} < 0{\cdot}12$ to be zero.

6) The data in the remaining dimensions with $\lambda_i^{1/2} > 0{\cdot}12$ are normalized according to the relation

$$z_{ij} = \frac{x_{ij}}{\lambda_i^{1/2}}\,, \qquad (6)$$

the class being then, in a geometrical approximation, a hypercube with an edge length (standard deviation) equal to unity.

7) The centre C_q for each class is calculated.

8) The mean of the squared distances $D_q^{(C)}$ of all prototypes of the class q from the centre C_q is estimated according to

$$D_q^{(C)} = \frac{\sum_{j=1}^{N_q} \sum_{i=1}^{R} (z_{ij} - c_i)^2}{N_q}, \tag{7}$$

where c_i is the mean of the dimension i.

9) The object to be classified (p) is autoscaled, transformed and normalized to the object t before the classification for each class separately by the same manner as y in the steps 1, 4 and 6 according to

$$t_r = \frac{\sum_{i=1}^{R} p_i' \cdot e_{ri}}{\lambda_r^{1/2}}, \tag{8a}$$

where

$$p_i' = \frac{p_i - \bar{y}_i}{(\overline{y_i^2} - \bar{y}^2)^{1/2}} \tag{8b}$$

and where t_r represents the r-th component of the object in the transformed space of class q.

10) The similarity of the object p to the corresponding class q is calculated as averaged squared distances of the object from all prototypes z in this class

$$D_q = \frac{\sum_{j=1}^{N_q} \sum_{i=1}^{R} (z_{ij} - t_i)^2}{N_q}. \tag{9a}$$

Alternatively, this similarity may be calculated on the basis of Eq. (*9b*) which is more convenient for computation purposes.

$$D_q = D_q^{(C)} + \sum_{i=1}^{R} (c_i - t_i)^2. \tag{9b}$$

11) For comparability, D_q are divided by the corresponding $D_q^{(C)}$

$$D_q/D_q^{(C)} = \frac{D_q^{(C)} + \sum_{i=1}^{R} (c_i - t_i)^2}{D_q^{(C)}} = 1 + \frac{\sum_{i=1}^{R} (c_i - t_i)^2}{D_q^{(C)}}. \tag{10}$$

Evidently, $D_q^{(C)}$ is constant for a given class. The first term, being equal to unity, can be neglected in the following comparative computation. Hencefore, the measure of similarity can be expressed in the form

$$S_q = \frac{\sum_{i=1}^{R} (c_i - t_i)^2}{D_q^{(C)}} . \tag{11}$$

Thus the measuie of similarity S_q is the ratio of squared distance of the object from the centre of a class and the averaged squared distances of this centre from all prototypes in the class.

12) For our dichotomy case, the object p is classified into class q_1 if $S_{q_1} < S_{q_2}$ and into the class q_2 if $S_{q_1} > S_{q_2}$. Generally, the object in the system with Q classes is classified into class q with the lowest value of S_q in comparison with those for Q-1 remaining classes.

13) The relative similarity S^+ of the classified object to the classes of the system is expressed as the ratio

$$S^+ = S_{q'}/S_q \tag{12}$$

of the measure of similarity for less similar class $S_{q'}$ and of that for more similar class S_q. For ratio (*12*) the relation $S^+ \geqq 1$ is valid. For $S^+ = 1$ no decision is possible. In general, the higher the S^+ the easier the decision.

The classification was checked by two measures of the recognition performance: the measure of correctness and the measure of reliability. The measure of correctness is the percentage of correctly recognized prototypes. The measure of reliability, M^r, is the ratio of summarized differences of similarities for correctly and incorrectly recognized prototypes

$$M^r = \sum_{i=1}^{n} (S_i^* - S_i) / \sum_{j=1}^{m} (S_j^* - S_j) ,$$

where n is the number of correctly and m that of incorrectly recognized prototypes. S^* is the similarity to "false" class(es) and S the similarity to "own" class.

EXPERIMENTAL

The data for 224 complex hydrides are identical with those published in ref.[11] and used in ref.[14,15]. Programs are written in the GIER-ALGOL-III version of ALGOL-60. The computations were carried out in the Computer Centre of the Institute of Nuclear Research, Řež.

RESULTS AND DISCUSSION

In Table I the comparative results of the classification at different dimensionalities of the MODEL-1 are given in percentage of accordance.

The Direct Comparison with the SIMCA *Method*

The principal task of the present work was testing our new method by means of the very efficient SIMCA method[12,13] using the identical data base[11]. We used the "classical" approach which measures the similarity of an object to a class by means of its average distance from all prototypes of the training set previously transformed and normalized by suitable preprocessing. The second, SIMCA approach used earlier[14] measures the similarity of an object by the degree of fittingness to the functional representation of the individual class, the function being of the principal components type. At the directly compared dimension 28 the measure of correctness for recognition is in the former case 89% whereas in the latter one 75% (ref.[14]). However, for purpose of direct comparison we discuss the percentage of the classification accordance (Table I). In 80% of the cases the "incorrectly" recognized hydrides are identical. This is very encouraging in the present study where rather different classification methods are compared. The results of the prediction of stability for 109 complex hydrides of "unknown" stability are identical to a significantly lower degree (68%) as can be expected on the basis of generally known worse performance of the prediction in comparison with that of the recognition.

TABLE I

Percentual Accordance of the Classification Results Using the Complex Hydrides Stability MODEL-1

Hydrides	Dimensionalities				
	28[a]/28	28/23	23/11	11[b]/11	28[a]/11
Prototypes	80	97	91	87	84
Stable prototypes	80	98	92	98	82
Unstable prototypes	80	95	90	75	85
Classified hydrides	68	97	79	—	72

[a] Reference[14]; [b] verification (see text).

The Influence of the Dimensionality Reduction

The second column of Table I gives the percentage of accordance of the recognition as well as the prediction after the reduction of dimensionality from 28 to 23 by neglecting the dimensions corresponding to $\lambda = 0$ (ref.[15]). The accordance in both cases is 97%. The additional reduction of dimensions from 23 up to the intrinsic dimensionality of the MODEL-1 $D^m = 11$ (ref.[15]) by deleting the dimensions corresponding to the lowest λ's ≈ 0 retains unchanged 91% of "incorrectly" recognized hydrides and 79% of predicted ones as shown in the third column of Table I.

Verification of the Classification

Because of the remarkable shortage of the prototypes in the class of unstable hydrides, we verified the classification performance at $D^m = 11$ by the "leave-*n*-out" method with six 16-membered subsets selected from the stable hydrides and with ten 2-membered subsets picked up from the unstable hydrides. Thus six different learning subsets with 79 stable prototypes and ten subsets with 18 unstable prototypes were available for the training procedure. The fourth column of Table I shows that the accordance amounts to 87%; 98% for the stable hydrides and 75% for the unstable ones. The latter value demonstrates that the recognition is very sensitive to the composition of a scarcely populated class. Nevertheless, the above accordance is too high for the recognition to be trivial due to the shortage of prototypes.

Results of the Classification

The results of the classification at $D^m = 11$ are compared with the results of the SIMCA method published recently[14] (see the fifth column of Table I). The accordance is 84% for recognition and 72% for prediction. The most remarkable differencies in both classifications can be summarized as follows:

The present method favours, in general, the stability of lithium aluminium hydrides *cf.* of $LiAlH_4$ and $LiAlH_nD_{4-n}$ where $n = 1, 2, 3$ and D = alkyl or alkoxyl. Fluoro and chloroderivatives are classified as stable, namely $NaBHF_3$, $NaBH_2F_2$, $LiBHF_3$, $LiAlH_3Cl$, $LiAlHCl_3$ and $NaAlHCl_3$. Products of hydrolysis are more frequently classified as stable entities than by the earlier method[14] *e.g.* in the case of $LiBH_n\,.\,(OH)_{4-n}$ ($n = 1, 2, 3$), $NaBH_n(OH)_{4-n}$ ($n = 2, 3$) and $KBH(OH)_3$. Some of the sodium alkoxyaluminium hydrides $NaAlH(OR)_3$ are classified as stable compounds contrary to the earlier results[14]. The thio derivatives $LiBH_2(SH)_2$ and $LiBH(SCH_3)_3$ are recognized as the stable compounds. The hydrides $LiBH_3OR$ (R = CH_3, tert-C_4H_9) are classified as stable agents. Contrary to the previous results[14] the hydrides $NaAlH_2(OCH_2CH_2OR)_2$ (R = C_2H_5, iso-C_3H_7) are predicted to be stable compounds and $NaAlH[OC(CH_3)_2CH_2OCH_3]_3$ to be an unstable one. In contrast to ref.[14] $CsTlH_4$ is classified as an unstable compound by the present method.

From the above differences it is evident that (*a*) the present classification method favours to some extent the stability of the complex hydrides under study in comparison with the recently used method[14], (*b*) this shift is particularly significant for whole subsets of hydride derivatives. Finally, it is worth of emphasizing that 65% of complex hydrides are recognized identically in all seven studied cases using the SIMCA method[14] and the new method presented here. This fact is very encouraging if we consider that the results were obtained by rather different classification methods at very different dimensionalities (from 49 up to $D^m = 11$) and using different verifications.

CONCLUSIONS

The classification results of two different methods are compared. They were found to give comparable results as indicated by 80% accordance in recognition and a 68% accordance in prediction. It was also found that the reduction of dimensionality up to the model dimensionality causes only very mild change in the accordance. These results are encouraging and provide an additional evidence that pattern – recognition analysis may become a very powerful tool even in the treatment of such crude models as the MODEL-1. The discussion of the differences shows that the classification decision should be consulted with the results of more than a single method only. If some objects remain unclassified even after such careful treatment, the only way is to improve the model. This is also our case and further work on the MODEL-1 is in progress.

The authors would like to thank Professor C. F. Bender, University of California, U.S.A., Professor P. C. Jurs, Pennsylvania State University, U.S.A., and Professor S. Wold, Umeå University, Sweden, for their valuable comment to the manuscript of this paper. They also thank Dr K. Kuchynka, J. Heyrovský Institute of Physical Chemistry and Electrochemistry, Czechoslovak Academy of Sciences, Prague, and Dr K. Eckschlager, Institute of Inorganic Chemistry, Czechoslovak Academy of Sciences, Prague, for their permanent encouragement.

REFERENCES

1. Sebestyen G. S.: *Decision-Making Processes in Pattern Recognition.* McMillan, New York 1962.
2. Nilsson N. J.: *Learning Machine.* McGraw-Hill, New York 1965.
3. Fu K. S.: *Sequential Methods in Pattern Recognition and Machine Learning.* Academic Press, New York 1965.
4. Patrick E. A.: *Fundamentals of Pattern Recognition.* Prentice-Hall, Englewood Cliffs, New Jersey 1972.
5. Andrews H. C.: *Mathematical Techniques in Pattern Recognition.* Wiley-Interscience, New York 1972.
6. Verhagen C. J. D. M.: Pattern Recognition *7*, 109 (1975).
7. Rotter H., Varmuza K.: Org. Mass Spectrom. *10*, 874 (1975).
8. Isenhour T. L., Jurs P. C.: Anal. Chem. *43*, 20A (1971).

9. Isenhour T. L., Jurs P. C. in the book: *Computer Fundamentals for Chemists* (J. S. Mattson, H. B. Mark, jr, H. C. MacDonald, jr, Eds), p. 285. Dekker, New York 1973.
10. Jurs P. C., Isenhour T. L.: *Chemical Application of Pattern Recognition.* Wiley-Interscience, New York 1975.
11. Štrouf O., Wold S.: *Data Base I: Stability of Complex Hydrides.* MODEL-1. Umeå University 1977.
12. Wold S.: Pattern Recognition *8*, 127 (1976).
13. Wold S.: *Pattern Cognition and Recognition (Cluster Analysis) Based on Disjoint Principal Components Models.* Techn. Report No 357, March 1974, University of Wisconsin.
14. Štrouf O., Wold S.: Acta Chem. Scand. *A31*, 391 (1977).
15. Štrouf O., Fusek J.: This Journal *44*, 1370 (1979).