SIMPLEX SEARCH FOR MATHEMATICAL REPRESENTATION OF CHEMICAL CLASS STRUCTURE

Oldřich ŠTROUF and Jiří FUSEK

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

Received January 21st, 1980

A simplex algorithm useful in the search for the mathematical representation of the class structure is described. The use of this algorithm in combination with an appropriate pattern recognition classification method is discussed from the point of view of its application in chemistry.

The simplex procedure was proposed in 1947 by Dantzig[1] as a computational method of linear programming and it was mainly applied by economists. In the sixties, the efficient sequential simplex procedure[2] and its modification[3] were developed and then successfully used in many different optimization problems including those of experimental chemistry[4–10]. In 1977, the utility, efficacy and reliability of the simplicial methods as well as their robustness in a stochastic environment were additionally increased by the Super Modified Simplex (SMS) procedure[11]. The simplex procedure is also convenient for the adjusting of optimum coefficients of mathematical equations including those which approximate the behaviour of chemical systems[6]. Such an application has been recently described[12–13] also for the pattern recognition classification of chemical objects by means of learning machine method[15]. This "simplex pattern recognition" was successfully used for adjusting of the optimum set of coefficients (weight vector) of linear discriminant function even in the case of linearly inseparable classes of the objects. This "simplex pattern recognition" thus enables more general application of the conceptually simple learning machine method.

In this paper we describe a new use of the simplex approach in pattern recognition analysis. By means of this approach we search for an appropriate mathematical representation of the class structure, the class being defined as a set of similar objects, in our case of chemical ones. The structure of the class can thus be accounted for as a type of relations between the relevant variables (features) characterizing the objects of the class. The search for the mathematical representation is, naturally, the sounder the better is the classification of the objects into the classes. For classification of chemical objects, the pattern recognition was successfully applied in the last decade[16,17]. There ae patern recognition methods working during classification with the knowledge of the class(es) structure'(s) or without. The former case can be illustrated *e.g.* by the Wold's SIMCA method[18,19], the latter case by our pattern recognition classification method developed recently[20]. Our method measures the similarity of objects by means of the Euclidean distance in a transformed and normalized space without considering any structural regularities during the classification process. Therefore, after the objects were classified into the appropriate classes, a search for the mathematical representation of the structures of the classes should follow in order to generalize the behaviour of similar objects in the class.

In this introduction a simplex algorithm suitable for such a search is described in the form used in the following paper[21] for a pattern recognition analysis of catalytic activity of some transition metals in ethane hydrogenolysis. Methodologically, our simplex approach uses some advantages of the SMS-method in an automated manner with the possibility of the operator's interaction in decision steps of the algorithm.

CALCULATIONS

All computations were carried out on Hewlett–Packard 9825 computer. The program was made for automatic performation of the procedure with displaying of the best simplex response found during the given iteration cycle together with the best and second best responses of the last iterative step of this cycle. This enables to check the convergency rate of the automatic iteration process and affect it interactively *e.g.* by changing the expansion (contraction) coefficient. The operator's interaction is necessary in the step *A1* and *E* (the formulation of different initial vectors) and in the step *F* (the suggestion of an additional type of mathematical representation).

Simplex Procedure

Geometrically, the d-dimensional simplex (d-simplex) represents a convex polygon defined by $d + 1$ vertices. These vertices are linearly independent vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{d+1}$. The simplest case is a 2-dimensional simplex which can be represented by a triangle (Fig. 1). Three-dimensional simplex is evidently a tetrahedron in 3-dimensional space.

In optimization procedure, the dimensionality of the simplex equals the number of the variables v_i ($i = 1, 2, \ldots, d$) and the set of the vectors $\mathbf{v}_j$ ($j = 1, 2, \ldots, d + 1$) to the number of experiments. The points $\mathbf{v}^B$, $\mathbf{v}^S$ and $\mathbf{v}^W$ in Fig. 1 corresponding to the positional vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ thus represent the initial situation of the optimization procedure (the initial simplex S_0) for experiments with two variables. Nevertheless, the multivariate situations ($d \geqq 4$) beyond the graphical representation are very frequent in the modelling of real systems. In these cases the starting situation of the simplex optimization procedure can be given by the tabulated data v_{ij} only (Table I). The data v_{ij} for real systems have to fulfill certain specific constraints. Thus the constraint of nonnegativity ($v_{ij} \geqq 0$) was formulated[22] for the "classical" economic problems of optimum planning. For the optimization of physical and chemical experiments the constraints are naturally formed by the extreme values of the individual variables ($v_i^{\min} \leqq v_{ij} \leqq c_i^{\max}$). Contrarily, for the optimization of mathematical equations where v_{ij} are the coefficients of the equations no special constraints are necessary.

A) *Generation of initial simplex.* In our algorithm the initial simplex S_0 is generated automatically in the following way:

A1) Choose arbitrarily the initial vector $\mathbf{v}_1$.

A2) Derivative linearly independent vectors $\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_{d+1}$ of the simplex $\mathbf{S}_0$ so that:

$$v_{ij} = v_{i1} \quad \text{for} \quad i \neq j - 1$$

and

$$v_{ij} = v_{i1} + q \quad \text{for} \quad i = j - 1 \,,$$

where

$$q = \left(\sum_{i=1}^{d} v_{i1}^2 / d \right)^{1/2} . \, 10^{-1} \,.$$

The initial simplex is thus generated in a way similar to the generation in the "simplex pattern recognition"[1,2], but in the latter case q is an arbitrarily selected constant.

B) *Responses at the vertices of the simplex.* In the optimization of experiments the responses r_j in Table I are the results of the experiments $\mathbf{v}_j$. The optimization is carried out by the estima-

TABLE I

Initial Simplex S_0 in the Form of Tabulated Data

			Variable, v_i					Response r_j
Experiment, v_j		i	1	2	.	.	d	
	j	1	v_{11}	v_{21}	.	.	v_{d1}	r_1
		2	v_{12}	v_{22}	.	.	v_{d2}	r_2
		.	.	.	.	.		.
		$d + 1$	$v_{1(d+1)}$	$v_{2(d+1)}$		.	$v_{d(d+1)}$	r_{d+1}

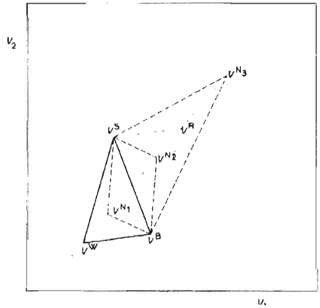

FIG. 1

Movement of Two-Dimensional Simplex

$\mathbf{v}^{Ni}$ is a new vector (vertex of the simplex) for different coefficients α : $i = 1$ for the contraction coefficient $-\alpha < 1$, $i = 2$ for the contraction coefficient $\alpha < 1$ and $i = 3$ for the expansion coefficient $\alpha > 1$.

tion of $\mathbf{v}_j$ with the maximum r_j in maximization problems and with the minimum r_j in the minimization ones. The optimization of mathematical equations is principally a procedure searching for the optimum set of coefficients of a given equation, *i.e.* the set which gives the closest fit of calculated and experimental values. The fit can be checked, *e.g.*, by least squares criterion[6] which is used also here for the optimization the of coefficients of the polynomial approximations.

B1) Estimate the responses r_j at the vertices $\mathbf{v}_j$ of the initial simplex $\mathbf{S}_0$ as the sums of squared differences between calculated values y'_{jk} and known values y_k, $k = 1, 2, \dots, K$, where K is the number of the known values

$$r_j = \sum_{k=1}^{K} (y'_{jk} - y_k)^2 .$$

E.g. for the linear polynom, the responses r_j are calculated according to:

$$r_j = \sum_{k=1}^{K} [\sum_{i=1}^{d} (x_{ki} v_{ij}) - y_k]^2 ,$$

where the values of the parameters $x_{k(d-1)}$ are estimated experimentally and the values x_{kd} are equal to one, the coefficient v_{dj} thus being an additive constant of the given linear polynomial equation.

B2) Order the vertices $\mathbf{v}^W$, $\mathbf{v}^B$ and $\mathbf{v}^S$ with the worst, the best and the second best responses, respectively.

C) *Moving of the simplex*. The simplest movement is the reflection[2] of $\mathbf{v}^W$ through the centre of the hyperplane (centroid $\mathbf{v}^C$), the hyperplane being defined by the remaining vertices after the deletion of $\mathbf{v}^W$. The result of this reflection is a reflected vertex $\mathbf{v}^R$. The reflection with expansion or contraction[3] represents the modification yielding a new vertex $\mathbf{v}^N$ instead of $\mathbf{v}^R$. The super modification[11] is ased also on the expanded (contracted) reflection, but in this case the location of $\mathbf{v}^N$ is estimated by means of the second degree curve constructed from the responses r^W, r^C and r^R. The analysis of the curves in all possible situations in maximization as well as minimization problems is discussed in the original paper[11]. In our minimization problem only that part of the analysis dealing with the "concavity up" case is used. In all remaining cases the simple expansion (contraction) scheme[3,6] of modified sequential simplex is followed.

C1) Delete $\mathbf{v}^W$ and calculate the centroid $\mathbf{v}^C$ from the remaining $\mathbf{v}_j$ according to:

$$\mathbf{v}^C = \sum_{j=1}^{d} \mathbf{v}_j / d .$$

C2) Calculate the response r^C at the centroid $\mathbf{v}^C$.

C3) Reflect $\mathbf{v}^W$ to obtain $\mathbf{v}^R$

$$\mathbf{v}^R = 2\mathbf{v}^C - \mathbf{v}^W .$$

C4) Calculate the response r^R at the vertex $\mathbf{v}^R$.

C5) Form a new vertex $\mathbf{v}^N$ according to the relation:

$$\mathbf{v}^N = \mathbf{v}^C + \alpha(\mathbf{v}^C - \mathbf{v}^W) ,$$

where α is an expansion (contraction) coefficient adjusted from the relations of responses r^W, r^B, r^S and r^C.

a) If $r^W + r^R - 2r^C > 0$, then relation between vectors and the corresponding responses is approximated by a parabolic function with a minimum (Fig. 2) (the "concave up" case in the SMS-procedure[11]). By the derivation of this function the expansion coefficient is then calculated as:

$$\alpha = (r^W - r^R)/2(r^W + r^R - 2r^C)\,.$$

b) Otherwise, if $r^W + r^R - 2r^C \leqq 0$, then the following scheme is used:

1) If $r^R < r^B$, then the expansion coefficient $\alpha = 1{\cdot}9$ is chosen.

2) If $r^S < r^R < r^W$, then the contraction coefficient $\alpha = 0{\cdot}5$ is used.

3) If $r^R < r^W$, then the negative contraction coefficient $\alpha = -0{\cdot}5$ is accepted.

4) Finally, if $r^B < r^R < r^S$ is valid, then no expansion (contraction) is carried out ($\alpha = 1$) and $\mathbf{v}^N \equiv \mathbf{v}^R$.

C6) Substitute $\mathbf{v}^w$ in the initial simplex $\mathbf{S}_0$ by the new vertex $\mathbf{v}^N$ forming thus the new initial simplex $\mathbf{S}_1$ and carry out the cycle *B1*—*C6* until following stopping criterion is fulfilled.

C7) Stop if the vectors $\mathbf{v}_j$ or/and the responses r_j do not change in at least two subsequent cycles. The resulting vector is then accounted for as the best one $\mathbf{v}^B$ of the initial simplex $\mathbf{S}_0$ generated from the first arbitrary $\mathbf{v}_1$.

D) *Simplex iteration.* After the cyclic procedure *A*—*C* is stopped according to the above criterion, the resulting $\mathbf{v}^B$ is accounted for automatically as a new initial vector $\mathbf{v}'_1$. The new initial simplex $\mathbf{S}'_0$ is generated from $\mathbf{v}'_1$ by the *A*-part of the algorithm and the search for the vector with the best response $\mathbf{v}^{B'}$ is carried out in the cycle *B*—*C*. This iteration process is made until the minimum sums of squared differences of r^B are practically identical in at least two subsequent iterations. The resulting vector is the best one for the simplices generated from the set of initial vectors being derived from the first initial arbitrary vector $\mathbf{v}_1$.

E) *Interactive search for true optimum.* To decrease the possibility of finding a false minimum, the procedure *A*—*D* is repeated with a new arbitrary vector $\mathbf{v}_1^*$ dramatically different from the

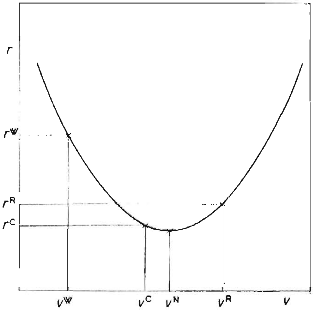

FIG. 2

Estimation of the Position of New Vector $\mathbf{v}^N$ in the Case of Responses Relation $r^W + r^R - - 2r^C > 0$ ("Concave up" case[11]) precedent initial vector $\mathbf{v}_1$. The best vector found by the procedure is compared with the best ones from previous simplex iterations. When mutually very different initial vectors give approximately the same values of the best responses r^B, then the corresponding $\mathbf{v}^{B*}$ is proposed to be a nearly optimum one for a given type of equation.

F) Search for appropriate type of equation. The coefficients of various tested mathematical equations (*e.g.* the polynomial equations of different degrees[21]) are optimized by means of the above procedure *A—E*. The the resulting best vectors of the tested equations are ordered according to the values of their responses r^B and, finally, the equation with the set of coefficients corresponding to the minimum best response is accounted for as the closest mathematical approximation among those treated in the study under consideration.

DISCUSSION

The principal task of chemistry is the extraction of information about the sought property (r in Table I) from the set of chemical data, here v_{ij} (Table I). The modelling is a very efficient tool for the analysis of chemical systems characterized by such data[23]. Many real chemical systems belong to multivariate systems with high value of d for v_i (Table I). Moreover, for numerous objects of the system the experiments (here $\mathbf{v}_j$) are cumulated by means of very efficient automated measurement equipments in a dramatically accelerating rate. For such complex chemical systems the modelling based on pattern recognition approach has been recently proposed[24]. This modelling consists of *a*) classification of the objects of the system into classes according to the similarity of objects with respect to a sought property, the similarity being estimated by an analysis of multivariate data[20], *b*) mathematical representation of the structure of the classes[21], *c*) estimation of the level of the sought property[25] and *d*) determination of intrinsic dimensionality[26] of the system by the dimensionality reduction and feature selection.

The simplex method described in this paper seems to be a useful computational method for the above parts *b*) and *c*), as shown by the modelling of the catalytic activity of transition metals in the hydrogenolysis of ethane[21,25]. The presented simplex method can be generally applied for the optimization of coefficients of different mathematical equations; in our modelling we tested polynomial equations[21,25] which are used also here for the simplex algorithm demonstration.

The main aim of the mathematical representation of chemical class structures is twofold: First, the information about structural homogeneity of the system, this information being necessary for the sound evaluation of the system behaviour. Secondly, the possibility of a rapid classification of new object with the sought unknown property by a very simple calculation.

REFERENCES

1. Dantzig G. B.: *Lineárné programovanie a jeho rozvoj*, p. 32. SVTL, Bratislava 1966.
2. Spendley W., Hext G. R., Himsworth F. R.: Technometrics *4*, 441 (1962).
3. Nelder J. A., Mead R.: Comput. J. *7*, 308 (1965).

4. Ernst R. R.: Rev. Sci. Instrum. *39*, 998 (1968).
5. Long D. E.: Anal. Chim. Acta *46*, 193 (1969).
6. Deming S. N., Morgan S. L.: Anal. Chem. *45*, 278A (1973).
7. Morgan S. L., Deming S. N.: Anal. Chem. *46*, 1170 (1974).
8. Parker L. R., jr, Morgan S. L., Deming S. N.: Appl. Spectrosc. *29*, 429 (1975).
9. Dean W. K., Heald K. J., Deming S. N.: Science *189* (4205), 805 (1975).
10. Johnson E. R., Mann C. K., Vickers T. J.: Appl. Spectrosc. *30*, 415 (1976).
11. Routh M. W., Swartz P. A., Denton M. B.: Anal. Chem. *49*, 1422 (1977).
12. Ritter G. L., Lowry S. R., Wilkins C. L., Isenhour T. L.: Anal. Chem. *47*, 1951 (1975).
13. Brunner T. R., Wilkins C. L., Lam T. F., Soltzberg L. J., Kaberline S. L.: Anal. Chem. *48*, 1146 (1976).
14. Lam T. F., Wilkins C. L., Brunner T. R., Soltzberg L. J., Kaberline S. L.: Anal. Chem. *48*, 1768 (1976).
15. Nilsson N. J.: *Learning Machine*. McGraw-Hill, New York 1965.
16. Jurs P. C., Isenhour T. L.: *Chemical Applications of Pattern Recognition.* Wiley-Interscience, New York 1975.
17. Eckschlager K., Štěpánek V.: *Information Theory as Applied to Chemival Analysis*, p. 155. Wiley-Interscience, New York 1979.
18. Wold S.: Pattern Recognition *8*, 127 (1976).
19. Wold S., Sjöström M. in the book: *Chemometrics. Theory and Application* (B. R. Kowalski, Ed.), p. 243. ACS Symposium Ser. 52, American Chemical Society, Washington 1977.
20. Fusek J., Štrouf O.: This Journal *44*, 1362 (1979).
21. Štrouf O., Fusek J., Kuchynka K.: This Journal, in press.
22. Dantzig G. B.: *Lineárné programovanie a jeho rozvoj*, p. 51. SVTL, Bratislava 1966.
23. Bykov G. V. in the book: *Modelirovanile v Teoreticheskoi Khimii* (Akad. Nauk SSSR, Ed.), p. 5, Izdatelstvo Nauka, Moscow 1975.
24. Štrouf O.: II Czech-Polish Colloq. on Chem. Thermodynamics and Phys. Org. Chem., Kazimierz 1980, Lectures, p. 103.
25. Kuchynka K., Fusek J., Štrouf O.: This Journal, in press.
26. Štrouf O., Fusek J.: This Journal *44*, 1370 (1979).

Translated by the author (O. Š.).