An Improved Optimization Strategy and Its Application to Clustering Analysis

Feng GAN,^{*†} Qingsong XU,^{**} Lin ZHANG,^{***} and Yizeng LIANG^{***}

**College of Chemistry and Biology, Jiangxi Normal University, Nanchang 330027, P. R. China*

***Department of Mathematics, Hunan University, Changsha 410082, P. R. China*

****College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China*

In this paper, a new optimization strategy is put forward which locates as many potential unimodal regions as possible in the search space. The potential optima can be further explored by a global optimization method for searching in the identified unimodal regions. The proposed strategy was evaluated by the optimization of test functions. The results obtained by this approach are comparable with those achieved by variable step size generalized simulated annealing (VSGSA) and a genetic algorithm (GA). Finally, we used this strategy in a clustering analysis of a tobacco data set.

(Received October 4, 2000; Accepted March 21, 2001)

Several global optimization methods, such as a genetic algorithm (GA) and simulated annealing (SA), have already been widely used in analytical chemistry.¹⁻⁴ Recently, a new global optimization method, sequential number-theoretic optimization (SNTO),⁵ has been developed in statistics. Some initial studies have been conducted to introduce this method into chemistry.^{6,7} SNTO is attractive due to its simplicity, ease of implementation and effective optimization performance.

The global optimization methods are superior to the classical optimization methods, such as simplex methods and the Newton method, because they can jump out from local optima. Although the classical methods can be implemented by running several optimization processes from different initial locations in the search space, it is still hard to guarantee that the algorithms will converge to the global optimum due to the fact that these methods essentially search locally.

However, one must be clear-headed concerning the existing global optimization methods, because they can reach the global optimum only when the number of runs is infinite. In a real optimization problem, convergence to the real global optimum can not be guaranteed, since only a finite number of runs can be performed in practice. Shaffer³ has presented a very nice discussion on GA and SA. The situation is just as he pointed out, because there is no "optimal" optimization method (Ref. 3, p. 241A, last sentence of paragraph 3). One should thus not be starry-eyed concerning the existing global optimization methods when facing a complicated chemical problem.

We think that a discrete optimization strategy is needed when applying the optimization methods to a real problem. Before we put forward our optimization strategy, let us reconsider the optimization process, which can be seen as an estimation to the global optimum. Thus the process is crucial for the final estimation. It has been proved that the so-called global optimization methods, such as GA, SA and SNTO, can reach the global optimum when the number of runs is infinite, that is

to say, make a real estimation to the global optimum. To a real problem with only one optimum, it is not difficult for the existing global optimization methods to make a real estimation of the optimum in limited runs. Unfortunately, because in a real problem there might be many local optima, the existing methods can not guarantee that one can make a real estimation to the global optimum in limited runs. It might be possible to estimate a local optimum, that is to say, be trapped into local optimization. This suggests that if one can obtain estimations of all the potential optima, it would be possible to obtain the real global optimum by comparing the estimations. To a real problem while one might not obtain all of the potential optima, as many potential optima as possible are needed. Because the global optimum is a maximal (or minimal), itself, it is generally ascendant for it to be selected into the potential optima set if one searches the potential optima in the same manner.

Based on the above discussion, we now put forward our optimization strategy. This strategy combines a number-theoretic net⁵ (NT-net) with the minimum spanning tree⁸ (MST) in graph theory to locate as many potential unimodal regions (which corresponds to potential optima) as possible. Once the unimodal regions are located, a global optimization method is applied to the regions for further searching. In this paper, SNTO is discussed for further searching, because it is easy to be implemented. References 5 and 6 list detail algorithms of SNTO. We, therefore, name our method graphic SNTO (GSNTO) for convenience of discussion.

Theory and Methodology

Number-theoretic net (NT-net)

An NT-net is a set of points which are uniformly scattered in an s -dimensional unit cube C^s . Several methods can be used to generate a uniformly scattered set in C^s . The NT-nets obtained by a good lattice point set (*glp* set) are often used in practice. In this research, a *glp* set is used. An NT-net $\{x_k = (x_{k1}, \dots, x_{ks}), k = 1, \dots, n\}$ can be easily generated by using the generating vector

[†] To whom correspondence should be addressed.
E-mail: ganfeng2001@sina.com

Table 1 Test functions

Name	Function	Domain
F1	$-2 \exp\{-0.5[(x-0.8)^2 + (y-2.6)^2]\} - 1.7 \exp\{-0.5[(x+2.5)^2 + (y+5.5)^2]\} - 1.9 \exp\{-0.5[(x-2)^2 + (y+2.7)^2]\}$	$-5 \leq x \leq 6, -6 \leq y \leq 7$
F2	$(4 - 2.1x^2 + x^4/3)x^2 + xy^2 + (-4 + 4y^2)y^2$	$-2 \leq x \leq 2, -1 \leq y \leq 1$

Fig. 1 Illustration of NT-net and minimum spanning tree (MST). o represents a good lattice point, which indicates the coordinates of the parameters. Point A and its nearest-neighbouring points can generate an MST.

$(n; h_1, h_2, \dots, h_s)$ in the *glp* set table and the following equation:⁵

$$q_{ki} = kh_i \pmod{n}, x_{ki} = \{(2q_{ki} - 1)/2n\}, \quad (1)$$

where $1 < h_i < n, i = 1, \dots, s, k=1, \dots, n; n$ is the number of points in an NT-net. For a more general domain, such as a rectangular region, $[a, b]$, NT-net $\{x_k = (x_{k1}, \dots, x_{ks}), k = 1, \dots, n\}$ can be obtained by following transformation:

$$x_{ki} = a_i + (b_i - a_i)y_{ki}, i = 1, \dots, s, \quad (2)$$

where, $\{y_k, k = 1, \dots, n\}$ is NT-net in C^s . Figure 1 shows an example of NT-net.

Appendix shows a routine for calculating the *glp* set.

Minimum spanning tree (MST)

In graph theory,⁸ a tree is an acyclic-connected graph. A minimum-spanning tree (MST) is a tree with minimum weight (generally speaking, an edge is a connection between two points and the weight is just the length of the edge). In this work, an MST is defined for every point in an NT-net as shown in Fig. 1. For a given point, an MST is centered at the point and embraces the nearest neighbouring points. An MST is labeled according to the name of the point generating the MST for convenience. For example, tree A is labeled by point A. Once a tree is constructed, the relationship of points can then be established. If the objective function value of point A is less than those of all other points in tree A, the tree A indicates a potential unimodal region. If point A does not have the above-mentioned feature, it remains unmarked.

Implementation details

Based on the above discussion, we propose the following algorithm:

Table 2 Position of potential local optima by our optimization strategy ($n_1 = n_2 = \dots = 233$)

Function	Potential optima	Obj. fun. val.	Coordinate (x, y)
F1	1	-1.6998	(-2.5130, -5.5036)
	2	-2.0000	(0.8000, 2.6000) ^a
	3	-1.8993	(2.0251, -2.7110)
F2	1	-0.2152	(-1.7031, 0.7916)
	2	2.1043	(-1.6094, -0.5683)
	3	-1.0316	(0.0901, -0.7141) ^a
	4	-1.0316	(-0.0906, 0.7127) ^a
	5	2.1044	(1.6120, 0.5697)
	6	-0.2154	(1.7045, -0.7949)

a. Global optimum.

samples No. 14 – 26 are cured tobacco.

All programs were written in MATLAB 4.2 and run in PC/586.

Results and Discussion

Optimization of test functions

Table 2 lists the results of our strategy for optimizing F1 and F2. From Table 2, one can see that our strategy can find the three potential optima of F1 and the six potential optima of F2. The real global optimum can then be obtained by comparing them. It must be noted that our strategy does not find all of the potential optima of F2. From the contour plot shown in Fig. 2, one can see that the distribution of the optima is symmetrical with eight optima. The reason that our strategy only finds six optima is that there are flat function value planes in the function. The two adjacent planes are marked as one unimodal region, which can be seen in Fig. 3. However, our strategy can find the two global optima, which are really important to real applications. The reason mentioned above that the global optimum is the maximal (or minimal), itself, is generally ascendant for it to be selected into the potential optima set if one searches the potential optima under the same situation. Thus, in general, our strategy will not miss the global optimum.

Table 3 lists the results by using our strategy and the original SNT0 to F1. One can see that the original SNT0 is trapped into a local optimum when the numbers of scattered points are $n_1 = n_2 = \dots = 144$. The reason is that in some cases the global optimum may lie far away from the best point obtained from the first uniformly scattered points in the studied space; thus the reliability of the results by SNT0 is lowered. This is not to say that such a situation will often be encountered by SNT0, though there seems to be no proper criterion to ensure that SNT0 makes a real estimation of the global optimum.

Table 4 gives the results of the GA, VSGSA⁴ and GSNT0 methods for the studied functions. The GSNT0 gives almost the same results as the GA and VSGSA method. It is worth noting that F2 has two global optima with almost the same objective function values around -1.0 at two symmetric coordinates in the searching domain. As one can see that the GSNT0 method can be used to determine two global optima of F2, while the GA and VSGSA methods can only determine out one of them, respectively. This feature might make GSNT0 a good method to deal with a complicated optimization problem, such as molecular modeling. For example, if there are two global optima with almost the same objective function values for a chemical system, one can conduct chemical experiments to

Fig. 3 Located potential unimodal regions by MST for the F2 function. o represents a good lattice point; * represents located potential unimodal regions.

further prove the results and to determine which one should be adopted. The computation time for all three methods are given in Table 5. It can be seen that the GSNT0 method is much quicker than the GA and VSGSA methods in this case.

Application to cluster analysis

Many clustering analysis methods have been introduced to chemistry.¹⁰⁻¹² One popular approach is to convert the problem of clustering into an optimization one, *i.e.*, minimizing or maximizing an objective function. Clustering with the best objective function value is the best one.

For a problem of clustering a set of N samples with their measurement vectors, $(x_1, x_2, \dots, x_N)$ into disjoint clusters $(C_1, C_2, \dots, C_k)$, the objective function can be written as

$$J_{se} = \sum_{k=1}^K \sum_{i=1}^{N_k} \|x_i^{(k)} - m_k\|^2 \quad (3)$$

where $x_i^{(k)}$ is the i th pattern belonging to cluster C_k and cluster C_k has N_k patterns; m_k is the mean of cluster C_k ,

$$m_k = \frac{\sum_{i=1}^{N_k} x_i^{(k)}}{N_k} \quad (4)$$

From Eq. (3), one can see that the Euclid distance has a complex nonlinear form, that there might be multiple optima, and that it is liable for an optimization method to be trapped into a local optimum. For example, although the iterative K-means algorithm developed by Coomans¹² *et al.* is an iterative procedure, it can not guarantee convergence to the global minima.

For every possible given cluster centers, say $m_1, m_2, \dots, m_k$, there is a corresponding J_{se} value. The clustering approach used in this paper is to find the best $m_1, m_2, \dots, m_k$ which give the minimal J_{se} value. Since the details of clustering theory have been extensively described in Refs. 12 and 13, readers who are interested in the details of the method can refer to them.

Table 6 gives the results of a clustering analysis of tobacco data based on our strategy. Two kinds of clustering are obtained. Figure 4 shows principal-component projection plots of the data; the two clustering results are shown in Fig. 4(a) and Fig. 4(b), respectively. In Fig. 4(a), the clustering result is about the same as that of Jin's with only two extra, which are samples 1 and 3. Samples 1 and 3 should be air-dried tobacco, but they

Table 3 Comparison of optimization results of F1 by our optimization strategy and SNT0

N_i	Our method		SNT0	
	Obj. fun. val.	Coordinate (x, y)	Obj. fun. val.	Coordinate (x, y)
$n_1 = n_2 = \dots = 144$	-1.9995	(0.7907, 2.5788)	-1.9000 ^a	(2.0000 ^a , -2.7000 ^a)
$n_1 = n_2 = \dots = 233$	-2.0000	(0.8000, 2.6000)	-2.0000	(0.8000, 2.6000)

Global optimum: Obj. fun. val. = -2.0000, $x = 0.8000$, $y = 2.6000$. a. Local optimum.

Table 4 Comparison of three optimization methods for optimization of F1 and F2

Function	GA		VSGSA		GSNT0	
	Objective Function value	Coordinate (x, y)	Objective Function value	Coordinate (x, y)	Objective Function value	Coordinate (x, y)
F1	-1.9997	0.7987 2.6021	-1.9998	0.7996 2.6002	-2.0000	0.8000 2.6000
F2	-1.0000	0.0000 0.7227	-1.0314	0.0963 -0.7149	-1.0316 -1.0316	-0.0903 0.7125 0.0895 -0.7114

GA: $N = 60$, $p_c = 0.2$, $p_m = 0.1$, $G = 0.8$, generation = 100. VSGSA: Initial step size is 10 percent of constrained domain. Step size increases or decreases by a factor of 1.2. Initial β is chosen to be 3.5 according to the probability requirement in Ref. 9. VSGSA terminates when step size is less than 0.0001. GSNT0: $n_1 = n_2 = \dots = 233$.

Table 5 Computation time (s)

	F1	F2
GA	241.38	242.44
VSGSA	1253.7	2329.9
GSNT0	49.82	113.37

Table 6 Clustering results of two kinds of tobacco by GSNT0

Objective function value	325.3	469.9
Cluster 1 (air-dried tobacco)	2, 4 - 13	4 - 11, 13
Cluster 2 (cured tobacco)	1, 3, 14 - 26	1 - 3, 12, 14 - 26

are clustered into cured tobacco. There are at least two things that generate this result. The first is that the clustering method is not good enough to produce correct clustering; the second is that the two samples are outliers of the air-dried tobacco. From Fig. 4(a), where these two samples are closer to the cured tobacco, and from the optimization function value, it might be reasonable to conclude that samples 1 and 3 are outliers of the air-dried tobacco. It is worth noting that Chen¹³ had made a robust diagnostic to the tobacco data; the conclusion is that samples 1 and 3 are really outliers of the air-dried tobacco. The fact is that because samples 1 and 3 are sunbaked yellow tobacco, the inner pattern is different from that of the air-dried tobacco. These results also show that because the 16 chemical components are good representatives of the inner pattern of the two kinds of tobacco, on the whole they can be clustered correctly.

From Table 6, one can also see that there is another kind of clustering in the system. Although Fig. 4(b) shows that it might be a possible clustering, the optimization function value tells us that it is a local optimum, and thus not good clustering. Because of the existence of a local optimum, it has room for an optimization method to be trapped into the local optimum. If one uses an optimization method in this problem and reaches this local optimum, it might be possible to conclude that the clustering method is not good or that the 16 chemical components of the tobacco are not a good representative of the inner pattern of the two kinds of tobacco. But, the reality is that

the optimization process is trapped into a local optima.

Conclusions

The proposed optimization strategy enhances the reliability of optimization by searching the potential optima. There are two things that guarantee this strategy to reach good results. The first is that the number-theoretic net is that because used to generate uniformly scattered points in the studied space, the searching are under the same situation. This enhances the chance of the global optimum to be selected in the potential optima set. The second is that the minimum spanning tree of graph theory will hold as many potential unimodal regions as possible for further searching in these regions. Because the searching is under the same situation, in general, it is ascendant for the global optimum be selected in the potential optima set.

One can also clarify that our strategy is a combination of both global searching and local detailed searching. Global searching ensures finding the potential optima, and local detailed searching makes the estimations of real optima more accurate.

Appendix

%function [GlpSet]=GlpSet (DimNum, PntNum, GenVect);
%This routine is used to calculate the good lattice point set

Fig. 4 (a) Objective function value, 325.3. + represents air-dried tobacco, o represents cured tobacco and * represents samples No.1 and 3. (b) Objective function value, 469.9. + represents air-dried tobacco and o represents cured tobacco.

%in interval (0,1)

%

% GlpSet calculated glpset

% DimNum dimension number, such as the number of parameters

% PntNum point number you need, such as the number of experiment to be implemented

% GenVect generating vector,

%

%Programed by YZ Liang 12/10/1995, Hong Kong Baptist University.

%Revised by Feng Gan 18/9/2000, Hunan University

%

%You can use this program in your work, but please do not change above notation

function [GlpSet]=GlpSet (DimNum, PntNum, GenVect)

GlpSet = zeros (DimNum, PntNum);

[m, n] = size (GenVect);

k = 0;

for i = 1:n

if GenVect(1,i) == PntNum

k = i;

break;

else

```

        k = 0;
    end
end

if k~ = 0
    for i = 1:DimNum
        for j = 1:PntNum
            GlpSet (i, j) = (j*GenVect (i + 1,k)-0.5)/GenVect
                (1,k)-fix((j*GenVect (i + 1,k)-...0.5)/GenVect(1,k));
        end
    end
else
    msgbox("Wrong point number!", "Error message", "warn");
    GenVect = [];
end
% end of routine

```

Generating vector table of two dimension

PntNum	8	13	21	34	55	89	144	233	377
vector1	1	1	1	1	1	1	1	1	1
vector2	5	8	13	21	34	55	89	144	233

Usage: Save above table as a matlab file, say GenVect. mat. When using, load the file and implement
[GlpSet] GlpSet(2,233,GenVect);

References

1. K. Hitchcock, J. H. Kalivas, and J. M. Sutter, *J. Chemometrics*, **1992**, 6, 85.
2. D. B. Hibbert, *Chemom. Intell. Lab. Syst.*, **1993**, 19, 277.
3. R. E. Shaffer and G. W. Small, *Anal. Chem.*, **1997**, 69, 236A.
4. J. M. Sutter and J. H. Kalivas, *Anal. Chem.*, **1991**, 63, 2386.
5. K. T. Fang and Y. Wang, "Number-theoretic Methods in Statistics", **1993**, Chapman & Hall, 105.
6. Y. Z. Liang and K. T. Fang, *Analyst*, **1996**, 121, 1025
7. L. Zhang, Y. Z. Liang, R. Q. Yu, and K. T. Fang, *J. Chemometrics*, **1997**, 11, 267.
8. G. Chartrand and L. Lesniak, "Graphs & Digraphs (Second Edition)", **1986**, Wadsworth & Brooks/Cole Advanced Books & Software, printed in U.S.A., 67.
9. W. B. Jin and Y. Dai, "Tobacco Chemistry", **1994**, Tsinghua University Press, Beijing, 50.
10. D. L. Massart and L. Kaufman, "The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis", **1990**, Chemical Industry Press, Beijing, 126
11. A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", **1988**, Prentice-Hall, Englewood Cliffs, New Jersey, 83
12. D. Coomans and D. L. Massart, *Anal. Chim. Acta*, **1982**, 136, 15
13. W. Chen, H. Cui, and Y. Z. Liang, *Anal. Lett.*, **1996**, 29, 1647