經營科學
第15卷 第2號
1998年 11月

Clustering by Accelerated Simulated Annealing

Bok Sik Yoon* · Sangbok Ree**

■ Abstract ■

Clustering or classification is a very fundamental task that may occur almost everywhere for the purpose of grouping. Optimal clustering is an example of very complicated combinatorial optimization problem and it is hard to develop a generally applicable optimal algorithm. In this paper we propose a general-purpose algorithm for the optimal clustering based on SA(simulated annealing). Among various iterative global optimization techniques imitating natural phenomena that have been proposed and utilized successfully for various combinatorial optimization problems, simulated annealing has its superiority because of its convergence property and simplicity. We first present a version of accelerated simulated annealing(ASA) and then we apply ASA to develop an efficient clustering algorithm. Application examples are also given.

I. Introduction

Clustering or classification is a very fundamental task that may occur almost everywhere for the purpose of grouping. Optimal clustering is one example of very complicated combinatorial optimization problems(NP-hard) and it is hard to develop a generally applicable optimal algorithm. In this paper, we tackle the clustering problem by an accelerated version of simulated annealing algorithm(ASA).

We consider clustering under quite general setting without any prior information on the number of clusters or the statistical structure of the given data. Statistical validation of the obtained clustering is not our direct concern here. Instead, we will try to find out most reasonable clustering based on the properly designed cost function.

Traditionally, there have been two different approaches in clustering: hierarchical methods and optimization methods. Among them, the hierarchical method has been used in most

* 홍익대학교 공과대학
** 서경대학교 산업공학과

practical clustering tasks mainly because of its simplicity[6, 7]. But the hierarchical method is just a heuristic approach and usually does not provide optimal solution. Some optimization methods have been proposed, mostly based on branch and bound techniques, but they failed to acquire their popularity mainly because of their enormous computational complexity[8].

In this paper we propose a two-stage method by combining the merits of two aforementioned approaches. In the first stage, we use a properly tuned hierarchical method to get a reasonably good clusters(, also the number of clusters) for the given data. In the second stage, we employ the ASA algorithm with the clustering result from the first stage as the initial solution. We include the intensification and diversification strategy of tabu search[3,4,5] to improve the efficiency of SA1 in Yoon and Cho(1996) further as will be explained in section 3. Since we take the hierarchical method as our starting point, our method will give a solution at least as good as one obtained by the hierarchical method. Moreover since ASA maintains the basic structure of the simulated annealing, the final solution is quite probably close to the optimal solution.

After this introduction, in section 2 we explain hierarchical clustering methods and cost functions briefly. In section 3, we give our ASA algorithm and it implementation schemes. After applying ASA to a set of qualitative data, we present the computation results in section 4, we finish our paper with a conclusion in section 5.

Ⅱ. Hierarchical clustering and objective function

In this study, we consider clustering under general situation. The data may be qualitative or quantitative and the number of clusters need not be specified a priori. The distance(similarity) measure is assumed to be given appropriately.

Clustering methods can be classified into two types: hierarchical methods and optimization methods(Hand, 1981). The hierarchical method can be done agglomeratively by combining the nearest neighbours at each hierarchical step until the proper number of clusters are obtained, or divisively by splitting a cluster into smaller subclusters at each step beginning with the whole data as a single cluster. The optimization method try to find the best clustering among all the possible combinations of clusters without any hierarchical steps.

Traditionally, the hierarchical method has been used with more popularity than the other mainly because of its simplicity. But in the (agglomerative) hierarchical clustering, once a point is assigned to a cluster at some level, it cannot be transferred to another cluster at the higher level. Thus in principle, the optimization method can provide better clustering since it considers every possible combinations of clusters. But in actual applications the optimization problem is solved approximately by branch and bound techniques or by some other heuristic algorithms[7, 8].

Another problem in the optimization method is to find a proper optimization criteria(i.e. the objective function). We try to find a clustering in which points within a cluster have close similarity and the different clusters have clear dissimilarities. In case of quantitative data obtained in R^m, letting $X^n = \{ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \}$ denote the set of n data points and $X_i = \{ \boldsymbol{x}_j | \boldsymbol{x}_j \in cluster\, i \}$, n_i = the number of points in cluster i, $\bar{\boldsymbol{x}}_i = \sum_{\boldsymbol{x} \in X_i} \boldsymbol{x}/n_i$, $\bar{\boldsymbol{x}} = \sum_{\boldsymbol{x} \in X^n} \boldsymbol{x}/n$, and c be the number of clusters, the total scatter matrix T, within-class scatter

matrix W and between class scatter matrix B can be defined as

$$T = \sum_{x \in X} (x - \bar{x})(x - \bar{x})^T, \quad (1)$$

$$W = \sum_{i=1}^{c} \sum_{x \in X_i} (x - \bar{x}_i)(x - \bar{x}_i)^T, \quad (2)$$

$$B = \sum_{i=1}^{c} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T, \quad (3)$$

respectively. Note that $T = W + B$, which implies maximizing W is equivalent to minimizing B. For the optimization purpose, we need to summarize W and B. Usually trace of W, determinant of W, trace of $W^{-1}B$, or trace of $T^{-1}W$ used for this purpose. For the case of qualitative data, an example of the optimization criteria will be discussed in section 4.

Ⅲ. Algorithm ASA

3.1 general form

We take ASA algorithm shown in figure 1 as our basic algorithm. ASA has been proposed and tested empirically in Yoon and Cho(1996). It is faster than standard simulated annealing (SA) and yet maintains the convergence property because it does not alter basic SA structure. The main feature of ASA is that it increases the speed of convergence by making the size of inner loop smaller than SA. This is made possible by checking if the Markov chain describing the evolution made in the inner loop of SA reaches the steady state and decreasing the temperature T only when the Markov chain is judged to reach the steady state. Of course there are other considerations adapted to make ASA more efficient: the automatic initial temperature setting mechanism, careful stopping criteria to avoid unnecessarily long wandering period at the final stage of the convergence, etc. The ASA algorithm is given in figure 1.

```
Algorithm ASA
INITIALIZE(X,T,L);
X_best = X;
Counter1 = 0;
Counter2 = 0;
repeat
    Costold = C(X);
    Check = 0;
    for i=1 to L do
      Y=PERTURB(X);
      if  (C(Y) ≤ C(X)) or
       (exp ((C(X) - C(Y))/ T) > random (0,1))
      then
         X=Y ; {accept the movement}
      endif;
      if (C(X) < C(X_best)) then
           X_best = X;{update the best solution}
           Counter2 = 0;
           Check = 1;{if 1, update T}
      else
           Counter2=Counter2 + 1;{for stopping}
      endif;
    endfor;
    Costnew = C(X);
    UPDATE (T, Costnew, Costold, Check);
    if(Costnew = Costold) then
        Counter1 = Counter1 +1;{for stopping}
    else
        Counter1 = 0;
    endif;
until (Counter1 > M or Counter2 >N);

UPDATE(T, Costnew, Costold, Check)
   if(Check=1 or Costnew < Costold) then
        T = aT;
   endif;
```

Figure 1. ASA Algorithm

3.2 Implementation schemes

To apply ASA for clustering problems, we need to specify the solution space and neighbourhood structure, the initial solution, the perturbation scheme, and the cost function.

If there are c clusters, a solution will be X_j, $j=1,\ldots,c$ with the notation defined in section 2.

For the initial solution we use the following step.

[initial solution]

Perform the agglomerative hierarchical clustering two times: once by the nearest neighbour(single link) method and once by the farthest neighbour(complete link) method. Set the one which gives better cost as the initial solution for ASA. Let the number of clusters of the initial solution be c.

The perturbation scheme involves two type of movements: near movements and far movements. In our perturbation scheme, near movements which is similar to the intensification step in tabu search[5] occur very frequently and far movements which is similar to the diversification step in tabu search occur rarely.

[near movement steps occurring with probability a(T)]

1. Choose one element u randomly from {1,..., n}
2. Generate a random integer I in {1, ..., c+1}.
   If I=c+1, make c=c+1.
3. Move u into the cluster I.

[far movement steps occurring with probability 1-a(T)]

Merge the two clusters with the minimum between-distance. Make c=c-1.

Note that the perturbation scheme makes it possible for the number of clusters to be increased or decreased, so that proper number of clusters can be obtained automatically during progress of the algorithm.

Ⅳ. Application examples

4.1 A qualitative data

We test our algorithm on a real world problem instance which is taken partially from Domdorf and Pesch(1994). We want to classify 50 lecturers according to their membership in commissions and academic interests. As we can see in table 1, there are 10 classification attributes.

With the data given in table I, we make a 50x10 data matrix $D=(d_{ik})$, where $d_{ik}=1$ if lecturer i has either a membership of a commission or interest in its subject and $d_{ik}=0$ otherwise, for $i=1,2,\ldots50$, $k=1,\ldots10$. With D, we define a measure of similarity between distinct objects(data points) as

$$w_{ij}=2|\{k|d_{ik}=d_{jk},\ k=1,\ldots,10\}|-10,\ 1\le i<j\le 50,$$

where |A| implies the cardinality of a set A, and the objective function as

$$\sum_{1\le k\le c}\ \sum_{i,j\in X_k}\frac{w_{ij}}{n_k^2},$$

which implies roughly the sum of mean similarity between all pairs of objects in each cluster. Of course, we try to find a clustering maximizing the objective function.

The algorithm is coded in C and run on a PC(IBM 486 DX4) under MS-DOS. In ASA we set L=50, a=0.95, N=M=5, and T is generated adaptively as explained in Yoon and Cho(1996). The initial feasible solution is generated as explained in section 3.2 and the objective func-

〈Table 1〉 Lecture example

Classification of Lecturers

The association of Business Administration Lectures consists of 5 commissions. Each "1" entry of the data matrix expresses either membership of a commission or interest in its subject. In the listing of the data below for each object (i.e., lecturer) we only grve the numbers of those columns corresponding to attributes with a "1" entry in the 0,1 data matrix. The number of the object is always printed in bold face.

Specification of all 10 Attributes

column	attribute
1/2	Member of the commission/Interest in International Management
3/4	Member of the commission/Interest in Information Management
5/6	Member of the commission/Interest in Operations Research
7/8	Member of the commission/Interest in Production Management
9/10	Member of the commission/Interest in Marketing

Data

1, 1 2 9 10; **2,** 3 4 6; **3,** 5 6 8; **4,** 1 2; **5,** 6; **6,** 7 8; **7,** 3 4; **8,** 1 3 9; **9,** 5 6; **10,** 6 8 9; **11,** 3 4; **12,** 3 4 5 6 7 8; **13,** 4 8; **14,** 2 9 10; **15,** 4 6 8; **16,** 5 6 7 8; **17,** 3 4 9 10; **18,** 1 3 5 7 9; **19,** 4 5 6 7 8; **20,** 2 4 6; **21,** 6 8; **22,** 10; **23,** 2; **24,** 3 5 7 9; **25,** 9; **26,** 6; **27,** 4 6; **28,** 4 6 8; **29,** 3 4 6 8; **30,** 7 8; **31,** 5 6 7 8; **32,** 8; **33,** 3 5; **34,** 4 5 6; **35,** 4 8; **36,** 5 6 7 8; **37,** 6; **38,** 7 8; **39,** 1 2; **40,** 9; **41,** 6 7 9; **42,** 9 10; **43,** 4; **44,** 6; **45,** 6 7 8; **46,** 3 4 5 6 9 10; **47,** 9; **48,** 9; **49,** 1 2; **50,** 6 8;

tion value of initial solution was 25.24. The best result obtained after running ASA 50 times is shown in Table 2. Average running time was 4.8sec and the best objective function value was 25.52. Even though the improvement in the objective function value was not so big, our method is still valuable because it could improve the initial solution which was the best solution obtainable from the hierarchical clustering method at the cost of small amount of additional computational efforts.

〈Table 2〉 Best solution

group number	number of data in each group	data number
1	18	1 14 22 42 4 39 49 5 26 37 44 8 25 40 47 48 23 32
2	5	7 11 13 35 43
3	6	3 9 16 31 36 41
4	4	6 30 38 45
5	3	10 21 50
6	2	12 19
7	3	15 28 29
8	2	17 46
9	3	18 24 33
10	4	20 27 34 2
Run time : 5.22 (sec) Objective function value: 25.52		

4.2 A quantitative data

We test our method with another data taken from Manly(1994). The data shows percentage of people employed in nine industry sectors of 26 countries of Europe. In Table 3, we can see the standardized version of data which was transformed from the original data to make each country's data have a mean of 0 and a standard deviation of 1. With the standardized data, we compute the Euclidean distances between all

pairs of countries and apply our method in section 3.2. In this example, we choose $\frac{B}{2c}$ as the objective function to maximize. Our objective here is to achieve a big value of B and at the same time to keep the number of clusters as small as possible. To determine a proper values of c in the initial solution step, we compute the objective function values for each c=1,2,...26 and find the c which gives the minimum objective values. The clustering results in the initial solution process appears in Table 4 and 5. Since the complete link method gives the bigger objective value, we take the resulting clustering as the initial solution and begin the ASA steps. The final result shows apparent increase in the objective value(Table 4). This seems to be very promising result because we could improve the best result of the hierarchical method with very small additional computation time(small fraction of total run time of 2.42 seconds).

〈Table 4〉 Nearest neighbor method

group number	number of data in each group	data number
1	17	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
2	1	18
3	7	19 23 20 21 22 24 25
4	1	26
Objective function value: 102.6		

〈Table 3〉 Transformed data

Country Number	Agr	Min	Man	Ps	Con	Ser	Fin	Sps	Tc
1	-1.02	-0.36	0.09	-0.02	0.02	1.34	0.78	0.96	0.47
2	-0.64	-1.19	-0.74	-0.81	0.08	0.36	0.89	1.78	0.40
3	-0.54	-0.47	0.07	-0.02	0.45	0.84	0.71	0.38	-0.61
4	-0.80	0.05	1.25	-0.02	-0.53	0.32	0.36	0.33	-0.32
5	0.26	-0.26	-0.90	1.04	-0.40	0.84	-0.43	0.11	-0.32
6	-0.21	-0.67	0.08	-1.08	1.12	1.12	-0.85	0.01	-0.61
7	-0.73	1.90	0.54	-0.29	0.63	1.21	0.21	-0.12	-0.25
8	-0.82	-1.19	-0.64	0.24	1.05	1.10	1.00	1.24	0.18
9	-1.06	0.15	0.46	1.31	-0.77	0.86	0.61	1.21	-0.10
10	-0.41	-0.16	0.46	1.31	0.51	0.84	0.32	-0.47	0.33
11	-0.39	-0.88	-0.16	1.04	-0.46	0.38	0.53	0.63	0.76
12	1.43	-0.67	-1.34	-0.82	-0.04	-0.32	-0.57	-1.32	0.11
13	-0.65	-0.78	-0.66	-0.29	0.26	0.86	0.25	1.11	2.05
14	0.56	-0.98	-0.36	-0.82	0.14	0.07	-0.46	-0.49	-0.61
15	0.24	-0.47	0.21	-0.55	2.03	-0.71	1.60	-1.20	-0.75
16	-0.84	-0.88	-0.16	-0.29	-0.59	0.31	0.71	1.81	0.18
17	-0.73	-1.09	1.54	-0.29	0.81	0.99	0.46	-0.68	-0.61
18	3.07	-0.57	-2.73	-2.15	-3.26	-1.70	-1.03	-1.19	-2.40
19	0.29	0.67	0.76	-0.82	-0.16	-1.08	-1.18	-0.27	0.11
20	0.17	1.70	1.21	0.78	0.32	-0.82	-1.11	-0.31	0.33
21	-0.96	1.70	2.03	1.04	-0.34	-0.38	-1.00	0.30	1.33
22	0.16	1.90	0.37	2.64	0.02	-0.78	-1.11	-0.41	1.04
23	0.77	1.28	-0.19	-0.02	0.14	-1.19	-1.11	-0.57	0.25
24	1.00	0.87	0.44	-0.82	0.32	-1.54	-0.96	-1.22	-1.11
25	0.29	0.15	-0.17	-0.82	0.63	-1.50	-1.25	0.52	1.98
26	1.90	0.25	-1.46	0.51	-1.98	-1.43	2.60	-2.16	-1.83

Agr=agriculture, Min=mining, Man=manufacturing, Ps=power supplies Con=construction, Ser=service industries, Fin=finances, Sps=social and personal services, Tc=transport and communications

〈Table 5〉 Farthest neighbor method

group number	number of data in each group	data number
1	12	1 3 4 9 11 7 5 10 2 16 8 13
2	9	6 17 15 12 14 19 23 24 25
3	2	18 26
4	3	20 21 22
Objective function value: 124.2		

〈Table 6〉 ASA final solution

group number	number of data in each group	data number
1	5	2 24 18 22 26
2	8	20 23 19 21 10 14 25 12
3	7	16 8 13 11 5 1 9
4	6	3 6 4 17 15 7
Run time: 2.42 (sec) Objective function value: 130.9		

V. Conclusion

In this paper a simulated annealing based algorithm for general clustering problems was proposed and its performance was tested with a qualitative data and a quantitative data. Because of the probabilistic convergence to optimality of ASA, we expect the algorithm gives a near -optimal clustering at least better than those can be obtainable by hierarchical methods. Moreover the algorithm is relatively easy to implement. We may further improve the performance of the algorithm by including some memory mechanism used in tabu search[2] or other stochastic evolution algorithms[9].

REFERENCES

[1] U. Dorndorf and E. Pesch, "Fast Clustering Algorithms," *ORSA Journal on Computing*, Vol.6, No.2(1994), pp.141-153.

[2] B. L. Fox, "Integrating and accelerating tabu search, simulated annealing and genetic algorithm," *Annals of Operations Research*, Vol.41(1993), pp.47-67.

[3] F. Glover, "Tabu search-Part I," *ORSA Journal on Computing*, Vol.1(1989), pp.190-206.

[4] F. Glover, "Tabu search-Part II," *ORSA Journal on Computing*, Vol.2(1990), pp.4-32.

[5] F. Glover, E. Taillard and D. D. Werra, "A user's guide to tabu search," *Annals of Operations Research*, Vol.41(1993), pp.3-28.

[6] D. J. Hand, *Discrimination and Classification*, John Wiley & Sons, New York, 1981.

[7] J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, New York, 1974.

[8] W.L.G. Koontz, P.M. Narendra and K. Fukunaga, "A branch and bound clustering algorithm," *IEEE Trans. on Computers*, Vol.c-24(1975), pp.908-915.

[9] J. Lee and M.Y. Choi, Optimization by Multicanonical Annealing and the Traveling Salesman Problem, working paper, (1994).

[10] B.F.J. Manly, *Multivariate Statistical Methods* (2nd ed.), Chapman & Hall, London, 1994.

[11] B. S. Yoon and G. Y. Cho, "Acceleration of Simulated Annealing ant Its Application for Virtual Path Management in ATM Networks," *Journal of the Korean Operation Research and Management Science Society*, Vol.20, No.2(1996), pp.125-140.