Bulletin
of
The Japanese Classification Society

Volume 4
December 1991
JAPAN

Contents

Development of a Graphical Interface for CART
Masaki Tsukamoto and Atsushi Ootaki

Generalization and Some Properties in Agglomerative Hierarchical Clustering Algorithms
Nagatomo Nakamura and Noboru Ohsumi

Similarity Coefficients for Classification from 3-way Binary Data
Shuichi Iwatsubo

Clustering Using a Simulated Annealing Algorithm
Yoshiharu Sato

Fitting Smooth Curves to Branching Data
Masahiro Mizuta

Measuring the Goodness-of-Fit of Logistic Models
Hideaki Hida, Takenobu Tasaki, and Masashi Goto

On the Criteria for the Acknowledgement of Minamata Disease
Masaya Miyai

Development of a Graphical Interface for CART

Masaki Tsukamoto and Atsushi Ootaki
Department of Precision Engineering
School of Science and Technology
Meiji University
Kawasaki Japan

CART (Classification And Regression Trees) by CalStat is a statistical software product for the analysis of classification and regression using the binary tree methodology. It would be convenient if the analysis results could be summarized graphically so that an analyst with a basic knowledge of statistics can consider the output from CART.

This paper describes the development of a graphical interface for CART which has the following functions.

(1) Binary tree sequences with split variables are plotted in analyses of both regression tre and classification trees.

(2) Graphical expression of regression tree analysis
It is not easy for an analyst with only a basic understanding of statistics to comprehend a data distribution simply through boxplots expression, although it is easy to find outliers. In this interface, data for a response variable at each terminal node are simultaneously shown in a histogram and boxplot together with those for the entire data. Also boxplots at all terminal nodes are shown for comparison with locations between arbitrary terminal nodes.

(3) Graphical expression of classification tree analysis
Correct classification cost and misclassification cost at each terminal node are shown in barplots for comparison with variation in misclassification cost between nodes. Also correct categories in misclassification at each terminal node are shown in separate barplots.

Generalization and Some Properties in Agglomerative Hierarchical Clustering Algorithms

Nagatomo Nakamura
4-12-20, Nodera, Niiza-shi
Saitama Japan

Noboru Ohsumi
The Institute of Statistical Mathematics
Tokyo Japan

Lance and Williams proposed the combinatorial solution, that is:

$$d_{(ij)k} = \alpha_i \, d_{ik} + \alpha_j \, d_{jk} + \beta \, d_{ij} + \gamma \, | \, d_{ik} - d_{jk} \, | \tag{1}$$

which d_{ij} indicates the dissimilarities between clusters C_i and C_j, and four parameters, α_i, α_j, β, and γ are used to characterize the clustering methods. For instance, they proposed the flexible method which is defined by $\alpha_i = \alpha_j = (1 - \beta)/2$, $\beta < 1$, and $\gamma = 0$ in expression (1).

Substituting the constants for the value of β (i.e., $\beta = -1/4$), we can obtain the conventional flexible method. However, substituting the flexible value for β in clustering process, we can find the new method which is based on the space-distorting properties. The new method lead to the result that single linkage and complete linkage methods are included in the flexible method as special cases. This indicates that it is possible to eliminate the parameter γ from expression (1). Furthermore, new parameter values are proposed that this method is given by specifying the parameters in expression (1) as follows:

$$\alpha_i = (1 - \beta) \bullet n_i / (n_i + n_j), \; \alpha_i = (1 - \beta) \bullet n_j / (n_i + n_j), \; \beta < 1, \; \gamma = 0.$$

where, n_i indicates cluster size of cluster C_i. The methods makes the advantage, we can see that include five methods in combinatorial solution (i.e., single linkage method, complete linkage method, group average method, weighted average method, and conventional flexible method).

For this situation, the method is named the *generalized flexible method.* The same method above, we can adjust the value of β using the space-distorting properties. For example, the method which keeps the space conservation, and which keeps the space dilatation. Anumerical experiment of the method is shown.

Similarity Coefficients for Classification from 3-way Binary Data

Shuichi Iwatsubo
Research Division
The National Center for
University Entrance Examination
Tokyo Japan

Some preliminary considerations are given. Similarity coefficients for 3-way binary data are introduced based on the simple matching coefficient, the coefficient of Jaccard and the eigen value problem of similarity matrix.

Clustering Using a Simulated Annealing Algorithm

Yoshiharu Sato
Department of Information Science
Faculty of Engineering
Hokkaido University
Sapporo Japan

Clustering n objects into k clusters under the within-class variance criterion can be viewed as a combinatorial optimization problem. Since the number of classifications is extremely large even for moderate n and k, exhaustive enumeration is not practical method. Most of clustering methods may be described as iterative techniques which improve the criterion successively. A well-known method of this technique is the k-means algorithm. Although these algorithms are computationally practicable and efficient, the solution inevitably converges on a local optimum neighboring the initial values.

As a method intended for a globally optimal solution, Koontz et al. (1975) proposed a clustering method based on the branch and bound algorithm of combinatorial optimization. However they pointed out that the naive application of this algorithm does not produced good results.

Recently, Kirkpatrick et al. (1983) introduced the concept of simulated annealing in combinatorial optimization. This can be used with any iterative improvement technique and may avoid becoming trapped in a local, rather than global, optimum.

In this paper, we develop a clustering method based on the simulated annealing algorithm and evaluate its performance.

The term "annealing" refers to a thermal process for obtaining a low energy state in a solid in condensed matter physics. The basic steps of this process are the following: Initially, maintain a very high temperature in the physical system, then decrease the temperature slowly until the system reaches the ground state (minimum energy). This annealing process can be modeled by using a computer simulation of the algorithm is known as the Metropolis algorithm (Metropolis et al., 1953) as follows: Given a current state i of a system with energy E_i, a subsequent state j is generated by a perturbation mechanism. The energy of the next state E_j. If $\Delta E = E_j - E_i \leqq 0$, then the state j is accepted as the current state, but if $\Delta E > 0$, then the state j is accepted with a probability

$$P_T(\Delta E) = \exp\{-\Delta E/kT\} ,$$

where T denotes the temperature and k a Boltzmann constant.

Kirkpatrick et al. applied this Metropolis algorithm to a combinatorial optimization assuming the following equivalences. States are equivalent to the solutions of the problem and the energy is equivalent to the cost of a solution. The temperature of the system is treated as a control parameter. Let f(i) and f(j) be the values of the cost function of states (solutions) i and j. The state transits from i to j if $f(j) \leqq f(i)$, but if $f(j) > f(i)$, the transition is accepted probabilistically according to

$$P_c(\text{accept } j) = \exp\{(f(i) - f(j)))/c\} .$$

In this algorithm, initially, at a large value of c, large deteriorations will be accepted. As c decrease, the acceptance of deteriorations become small; as the value of c approaches 0, no deterioration will be accepted at all. This feature plays an important role in obtaining the global optimum, because the possibility that the deteriorations can escape from the local optimum.

In a clustering problem, as the cost function we take the sum of the within-class variances. The state corresponds to the state of classification. Assuming that no objects are observed with respect to p variates,

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}),\ i=1,2,\ldots,n.$$

The Euclidian distances d_{ij} are viewed as the dissimilarities between the pairs of objects:

$$d_{ij} = \Sigma_r(x_{ir}-x_{jr})^2 \ .$$

Let the n × k matrix U represents the state of clustering, that is,

$$U = (u_{i\mathbf{a}}) \ , \ i=1,2,...,n; \ \mathbf{a}=1,2,...,k,$$
$$u_{i\mathbf{a}} = \ 1 \text{ (if the object i belongs to cluster a) ,}$$
$$= \ 0 \text{ (otherwise) .}$$

When n objects are divided into k clusters, the within-class variance of each cluster W_a and their sum W are denoted by

$$W_{\mathbf{a}} \ = \ \sum_{i<j}^{n} \ d_{ij}{}^2 u_{i\mathbf{a}} u_{j\mathbf{a}} / n_{\mathbf{a}} \ . \qquad n_{\mathbf{a}} \ = \ \sum_{i}^{n} \ u_{i\mathbf{a}} \ ,$$

$$W = \sum_{\mathbf{a}=1}^{k} \ W_{\mathbf{a}} \ .$$

The change in W when object i change its membership from cluster a to cluster b is denoted as

$$\Delta W = \sum_{i<j=t} d_{ij}{}^2 u_{i\mathbf{b}} u_{ij} / n_{\mathbf{b}}(n_{\mathbf{b}}+1) - \sum_{i<j=t} d_{ij}{}^2 u_{i\mathbf{a}} u_{j\mathbf{a}} / n_{\mathbf{a}}(n_{\mathbf{a}}-1)$$

$$+ \sum_{i=t} d_{it}{}^2 \{u_{i\mathbf{a}}/n_{\mathbf{a}} - u_{i\mathbf{b}}/n_{\mathbf{b}}+1)\} \ .$$

The state transition (change in membership) is accepted by the following probabilities forfixed control parameter c.

$$P_c = 1 \ , \qquad \text{if } \Delta W \leqq 0 \ ,$$
$$= \exp\{-\Delta W/c\} \ , \qquad \text{if } \Delta W > 0 \ .$$

Applying this algorithm to the data from two normal populations,

$$\mu_1 = (0.0, 0.0)', \qquad \Sigma_1 = I \ ,$$
$$\mu_2 = (1.5, 0.0)', \qquad \Sigma_2 = I \ .$$

For each population, 10 samples are generated with pseudo-random numbers, totaling 20 samples. In this numerical example example, the initial value of the control parameter c is given as 5.0, and the decreasing process is

$$C_{q+1} = 0.9 \ C_q \ , \ q = 1,2,....$$

The number of transitions m at q-th iteration is $m = n_k$. For k = 2 and k = 3, the results are shown in Table 1, where 100 initial partitions are given. For each initial partition, the clustering is performed and simultaneously k-means clustering is carried out.

It is a trivial result that, compared with the k-means method, the simulated annealing algorithm worked well, because the amount of the computation is much different. We should conduct a comparison the method by Koontz et al. who used the branch and bound algorithm. This is a problem for future consideration.

G. De Soete et al. applied the simulated annealing algorithm to uni-dimensional scaling. They pointed out that, contrary to theoretical expectations, there does not appear to be any overall advantage in using this algorithm. But our experiments or clustering show an advantage in the sense of attaining a globally optimal solution. The essential problem of simulated annealing seems to be the assignment of an initial temperature c and its cooling schedule, and the number of iterations for each temperature. There is trade-off between the number of iterations and the cooling schedule.

References

(1) De Soete, G. Hubert,L. & Arabie,P. (1988). The comparative performance of simulated annealing on two problems of combinatorial data analysis. *Data Analysis And Informatics, V,* E.Diday(Eds).p.489.

(2) Kirkpatric,S., Gelatt,Jr.C.G. & Vecchi, M.P.(1983). Optimization by simulated annealing. *Science 220,* p.671.

(3) Koontz,L.G., Narendra,P.M. & Fukunaga,K. (1975). A branch and bound clustering algorithm. *IEEE Trans. on Computers, C-24,* p.908.

(4) Metropolis,N., Rosenbluth,A., Rosenbluth,M., Teller,A. & Teller,E. (1953). Equation of state calculation for fast computing machines. *Journal of Chemical Physics, 6,* p.1087.

Table 1. Simulated annealing (SA) and k-means method (KM)
n=20, k=2,3; m=nk, c=5 (initial)

	SA		KM	
	W	Freq./100	W	Freq./100

k=2
optimum
W=26.63
	26.63	100	26.63	8

			27.36	35
			27.61	57

k=3
optimum
W=16.47
	16.47	100	16.47	5

			16.80	38
			16.81	39
			others	18

Fitting Smooth Curves to Branching Data

Masahiro Mizuta
Faculty of Engineering
Hokkaido University
Sapporo Japan

Yasumasa Baba
Department of Statistical Methodology
The Institute of Statistical Mathematics
Tokyo Japan

This paper presents a new method for fitting smooth curves to branching data. The fitting curves are constructed from three or more curves which connect at branching points.

Fitting smooth curves is one of the most important problems in data analysis. Simple regression analysis or multivariate regression analysis are used for data set consisting of observations on some variables, one of which can be treated as the response variable and the others as explanatory variables. However, these analyses do not work well for data sets whose variables cannot distinguish between response and explanatory.

Smooth curves can be fitted to data sets whose variables have been leveled using principal component analysis or generalized principal component analysis. In particular, Hastie and Stuetzle (1989) proposed the concept of principal curves and an algorithm for finding principal curves. Principal curves are nonparametric and their shape is suggested by the data.

In many situations, a data set cannot be fitted with a single curves, for example, a data set which spreads into two branches. We extend the concept of principal curves to includecurves with branching points. The algorithm is almost same as that for principal curves. Anumerical example of the proposed method is shown.

This study was carried out under the ISM Cooperative Research Program (90-ISM CRP-14).

Measuring the Goodness-of-Fit of Logistic Models

Hideaki Hida, Takenobu Tasaki, and Masashi Goto
Shionogi Kaiseki Center, 1-22-41, Izumicho
Suita City, Osaka 564, Japan

We examined the behaviors of several goodness-of-fit measures to evaluate the size of residuals in the fitting of logistic models, based on a re-analysis of the data from published literature and simulation study. The re-analysis confirmed that the measures LD (by Liu & Dyer, 1979) and G (by Green, 1998) were similar in their values. Their values were always larger than those of the measure E (by Efron, 1978). E was sensitive to outliers and all the measures were useful in determining the degree of polynomial models. The simulation study suggests that the values of all measures tended to become smaller as the errors became larger. It also suggests that the behavior of LD correlated best with those of the correlation coefficient or the proportion of variation explained by the model, and that the measure HGT (1/8) proposed by us had large variances and was unstable.

On the Criteria for the Acknowledgement of Minamata Disease

Masaya Miyai
College of Liberal Arts
Hieji-Dokkyo University
Hiemji Hyougo Japan

Minamata disease was formally detected on May 1, 1956, and was acknowledged as toxicosis by methyl mercury by the Japanese government in 1968. However, even today, not all of the problems have been solved yet.

We consider the criteria for acknowledging whether a patient suffers from Minamata disease or not using data derived from patients examined by Dr. Harada.

The clinical criteria defined by the Japanese Ministry of Welfare is that the patient must have some kind of combination of the specific clinical symptoms derived from the Hunter-Russel syndrome.

Linear discriminant analysis (quantification II) was applied to the data, using acknowledgement by postmortem examination as the objective variate and 17 clinical symptoms as the explanatory variates.

The results indicate that the effective combination of variates (symptoms) for discriminating whether a patient suffers from methyl mercury toxicosis is oral and peripheral sensory disorders. This is inconsistent with the accepted criteria. A medical paper written by Hunter and Russel reported that several specific symptoms are observed in individual patients disjointedly, not jointly.

Legal actions claiming acknowledgement as Minamata disease suffers were brought before the Kumamoto, Osaka, Kyoto and Tokyo District Courts by patients who were not acknowledged by the judging committee.

The Kumamoto District Court ruled that the clinical criteria were too strict to acknowledge the patients as Minamata disease sufferers and that the plaintiffs must be acknowledged.

Thus we came to believe that the clinical criteria are not consistent with the actual states and must be re-examined, and that the pathological criteria must also be re-examined.

In this case, discriminant analysis has little meaning. In addition, we think it is natural the data not be divided into two groups, acknowledged and unacknowledged, but one group, acknowledged.