Bulletin
of
The Japanese Classification Society

Volume 1
December 1987
JAPAN

Contents

Statistical Analysis of Heavy Metals
in
Human Body Organs

Masaya Miyai

Himeji-Dokkyo University
7-2-1, Kamiohno, Himeji, Hyogo, JAPAN

According to the industrial developments of Japan, substances which exist formerly few in environment increase gradually over the natural and human area and finally harmed human bodies.

In this report behaviours of mercury in human body organs are examined. Mercury, especially methyl, is one of the dangerous substances for all life over the world. It attacks the motor nerve, sensory nerve, optic nerve and so on, and causes complicate symptoms, called typically by Hunter-Russel's syndrome.

Data are measured on human dead body organs who thought to be suffered by mercury through environment originally exhausted from factory.

In this case organs means four parts, cerebrum, cerebelli, liber and kidney. Mercury means two kinds, total and methyl. Number of case is ninety eight constituted by seventy six males and twenty two females and have the other various kinds of characteristics each other, also they are divided into two groups, admitted to be harmed by mercury or not by the committee.

Each of eight variates distributes nearly log-normally and transformed to nearly normal distribution by logarithmic transformation.

Ratios methyl/total of each four organs distributes nearly log-normally and distributes wider range compared with former reports.

Almost all of ratios cerebrum/liver and cerebelli/liver of methyl mercury distribute within one. Half number of ratios of cerebrum/kidney and cerebelli/kidney of methyl mercury distribute over one.

Principal component analysis is applied to the data to see the structure of mercuries in human body organs.

The results show that the first principal component is thought to be the total amounts of each mercury, the second is the difference among organs, liver and kidney to cerebrum and cerebelli, the third is the difference between mercuries, methyl to total.

Each mark of admitted or not for ninety eight case is plotted on the sample score diagrams, but it is difficult to see the relation among the components and the groups of admitted or not.

Discriminant analysis is applied to the data to see whether the amounts of mercuries of each organ contributes to the decision of admittance, and if so then what kinds of mercuries do so.

The results shown that the most effective one is total merculy of kidney and next one is total merculy of cerebelli, but their contributions are not so great and each methyl mercury in question has few contributions.

Changes in Distribution of Accumulated Savings

Toji Makino
Takashi Hatada

Science University of Tokyo
Noda City, Chiba 278, Japan

Every autumn, the "Public Opinion Survey on Saving" conducted in the previous year by the Saving Encouragement Central Committee is published in most newspapers. We discuss here the distribution of accumulated savings in the light of this survey's results. It has already been shown that the distribution of savings, like income distribution, approximates to the log-normal distribution. In this presentation, we study the pattern of the distribution of accumulated savings over the last five years.

In the first place, we calculate the parameters and mode of the corresponding log-normal distribution from the mean and median of accumulated savings for the last five years. Further, we examine the change in the mean, median and mode of these saving holdings year by year.

Secondly, we draw Pareto diagrams for the distribution of accumulated savings in the last five years, and calculate the area of the bow shape that shows the degree of inequality. We examine the year by year variation in this.

Finally, we do ABC analysis on the Pareto diagram. We also examine the yearly variation in the points of division derived from ABC analysis.

From these calculated results, no great changes are evident in the savings differentials over the period of the study.

Disease Classification in Traditional Chinese Medicine

Hideo Miyahara and Han Jing Xian

**Dept. of Internal Medicine, School of Medicine, Kitasato University.
1-15-1, Kitasato, Sagamihara, Kanagawa, 228 Japan**

By using 28 signs and symptoms (SS) commonly used in traditional Chinese medicine, specialists in Chinese medicine classified 377 cases of chronic epigastralgia (wei won tong) into 3 disease categories, that is D1, D2 and D3. The data collected in this study were analysed by principal component analysis and Hayashi's quantification method type III. Both 377 cases and 28 SS were arranged according to the similarity and were assigned on multidimensional coordinates. On the 2 dimensional plane spanned by the first 2 components, 377 cases were plotted and were examined whether the members of the same category could make an isolated cluster and did not mix with those of the different categories. The examination revealed that the members of D1 and those of D2 occupied the left half and the right half of one unseparable group, respectively. On the other hand, those of D3 were scattered below the combined group of D1 and D2. As a result, it was concluded that the separation of D1 and D2 was considerably difficult from the standpoint of the numerical taxonomy alone, whereas that of D3 and the rest members was achieved by using the coordinates of the 2nd component. Spatial allocation of each of the 28 SS was represented multidimensionally, and could not be reduced a few dimensional space.

Automatic Thresholding Methods from Histograms

Nobuyuki Otsu

Electrotechnical Laboratory
1-1-4 Umezono, Tukuba-shi, Ibaraki, 305 JAPAN

Threshold selection is a basic problem in image processing and many other fields. It can be viewed as one-dimensional unsupervised classification, given a histogram as a mixture of implicit class distributions. Many methods have been proposed including adhocs and heuristics. In this paper, we briefly survey those methods and reconsider the problem from the standpoint of unsupervised statistical decision and also of information theory. Without loss of generality, two class cases are considered.

As a direct and desirable criterion function J(T) to evaluate the goodness of a threshold T and derive an automatic thresholding by optimization, we adopted the error rate of classification. The error rate would be minimized by the Bayesian thresholding if we knew exactly the two class distributions (supervised cases). However, in our case of unsupervised decision, we have to estimate the parameters and approximate the optimum decision in terms of threshold T.

Three methods are derived. One is based on the information theoretic equivocation as an upper bound on the Bayes error rate. The others make use of linear approximations of Bayes posteriors, for an upper bound and for the Bayesian thresholding in an iterative process, respectively. Some experimental results are shown to indicate the validity of the methods.

Writer Identification Experiment Using Reduced Variance Estimators

Isao Yoshimura

Faculty of Engineering, Nagoya University
Chikusa-ku, Nagoya, 464, JAPAN

Mitsu Yoshimura

Shotoku-Gakuen Women's Junior College
Nakauzura, Gifu, 500, Japan

The principal problem of writer identification is to infer the writer of a set of handwritten characters whose writer is not known from a reference set of characters whose writers are known. In this paper the following situation is considered:

'The character Y in question is ob served as a p-variate vector and the writer of Y is known to be among a persons Π_i,i=1, ... , a. The reference set is composed of the sample characters $\{X_{ij}$;i=1, ... ,a, j=1,...,$m\}$ in which the same letter as Y is written m-times repeatedly by each of the a persons. The writer of Y is inferred as the person the distance from which to Y is the minimum. The distance is assumed to have the form that

$$D^2(Y:\Pi_{i.}) = (Y-\overline{X}_{i.})'\hat{\Sigma}_i^{-1}(Y-\overline{X}_{i.}) + n^{1/2}\hat{\Sigma}_i^{1/2}, \qquad (1)$$

where $\overline{\mathrm{X}}$

is the sample mean of $\{X_{ij}$; i=1, ... ,$\mathbf{a}\}$ and $\hat{\Sigma}$

is a suitably chosen estimator of the variance matrix $\sum_i$.

Concerning with such situations Inaba [1] suggested the use of reduced variance estimator $W_i(w)$ defined by (2).

$$W_i(w) = (1-w)V_i + w\overline{V}. \qquad (2)$$

where V_i is the sample variance matrix based on

$$\{X_{ij};\ i=1,\ \dots,\ \mathbf{a}\} \text{ and } \overline{V}. = (V1+ \dots +V\mathbf{a})/\mathbf{a}.$$

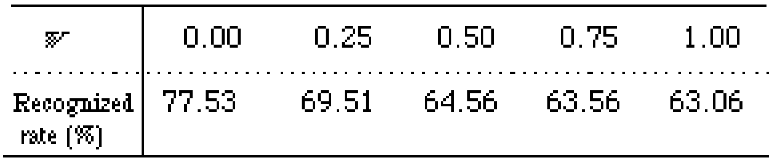
He showed the effectiveness of this estimator in cases of small m through a Monte Carlo study under normality assumptions. However in real situations the normality assumption is not always assured. The effectiveness must be examined through experiments with real data, which is performed in this paper.

The material used in the experiment is composed of repeated writings of a sentence with 25 letters by a=24 persons. The writings were done at 6 different occasions at least three weeks apart and at each occasion the sentence was written 8 times repeatedly. Two characters of each letter by each person at the first through the fourth occasions are contained in the reference set, that is, m=8. Four characters of each letter at each occasion by each person not included in the reference set are contained in the test set which is used for obtaining correct identification rates in the experiment.

As the features 'the weighted direction index histogram' proposed by Kurita et al. [2] are used which determine p=64 dimensional vectors. The parameter w is set on the five values 0.00, 0.25, 0.50, 0.75, and 1.00

In the calculation of inverse matrix of $W_i(w)$, eigenvalues less than the (m-1)-th are substituted with the (m-1)-th as is proposed in Yoshimura et al. [3].

The average correct recognition rates obtained in the experiment are as follows:



w	0.00	0.25	0.50	0.75	1.00
Recognized rate (%)	77.53	69.51	64.56	63.56	63.06

This result implies the superiority of the pooled variance estimator which corresponds to the reduced variance estimator with w=1.0. The result is out of the anticipation induced from Inaba's Monte Carlo study. The reason must exist in the following:

(a) The normality is violated.
(b) The population variances are equal.
(c) When m is less than p Inaba's result is not valid.
(d) When p and m are large Inaba's result is not valid.
(e) The substitution of eigenvalues yielded the result.

It is left for further study which is the true cause.

References

[1] Inaba. (1987), "Reduced variance estimators for discriminant analysis, "(in Japanese) 9-th symp. on applied statistics, 1987.10., Tokyo.
[2] Kurita et al. (1983), "Handwritten character recognition based on weighted direction index histogram and pseudo Maharanobis' distance," (in Japanese) IECE Jpn., PRL. 82-79.
[3] Yoshimura et al. (1986), "Discriminant analysis applied to the writer identification," 2nd Japan-China Symp. Statistics, 1986.11, Fukuoka, Japan.

Hierarchical Classification of Large Data Sets and its Applications

Miyuki Takahashi and Noboru Ohsumi

Department of Statistical Methodology
The Institute of Statistical Mathematics
Tokyo Japan

Various ideas will be required in order to classify large data sets under a computer environment, such as a microcomputer or workstation, and then watch the clustering process by graphical and colored representation.

The objective of this paper is to propose some procedures for presenting the skeleton of the data structure, and not to directly display the distribution of large-scale data. The clustering process for classifying large-scale data sets (exceeding 10,000 observations) and representing the classification results smoothly with computer graphics is provided. The process is essentially summarized as follows:

Step 1: A partial data set is sampled from a given data set at random in a certain proportion.
Step 2: Sampled data is subjected to the initial classification. At this time, the number of clusters is specified to be as large as possible to generate a large number of groups. The distance table is not used because it requires the use of a large matrix. The k-means method is used in a hierarchical manner to obtain the tree structure.
Step 3: The classification thereby completed is displayed by tree representation as the "preliminary classification."
Step 4: All individuals in the data set subjected to the initial classification areclassified consecutively using the tree obtained in Step 3 as a decision tree (i.e., each individual is assigned consecutively to some cluster in the tree by a binary decision rule).
Step 5: After all individuals are assigned, "reclassification" or "refinement" is conducted using the centroid vector of each cluster at the terminal end of the tree. The reclassification obtained in this step is the fine adjustment of the relationship between each cluster and the individuals belonging to each cluster. At this time, the tree structure is reconstructed using the RNN-rule (reciprocal nearest neighbor rule) andthe NN-chains (nearest neighbor chains).
Step 6: The tree structure obtained by reclassification is displayed and observedagain, and is compared with that obtained by the preliminary classification.

Next the experimental results obtained using artificial data sets and practical LANDSAT data are presented. The following figures show the tree structures presenting the results of preliminary classification and the reclassification of these data sets. In Figures 1 and 2, it can be seen that the figures of the original artificial data are fairly well exposed by the trees. Onthe other hand, in Figures 3 and 4, we can observe several tendencies indicating the characteristics of LANDSAT data, that is, the heterogeneous classes (the difference between land and sea,the sates of ocean currents and clouds, and so on) within the data sets can be clearly observed from the shape of branches in the tree structures. These features will be more really emphasized using the color display monitor and color photo prints.

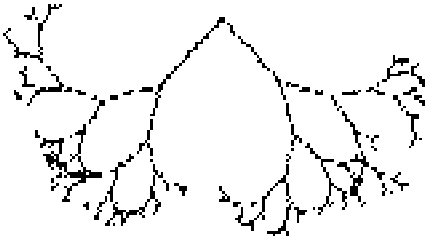


Figure 1
Result of preliminary classification
of artificial datasetof artificial dataset
(61,440 observations)

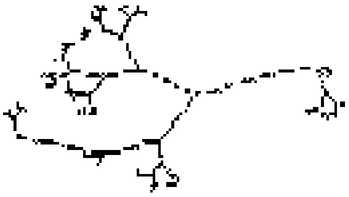


Figure 2
Result of reclassification
(61,440 observations)

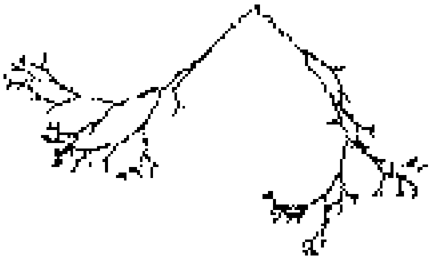


Figure 3
Result of preliminary classification
of LANDSAT data
(61,440 observations)

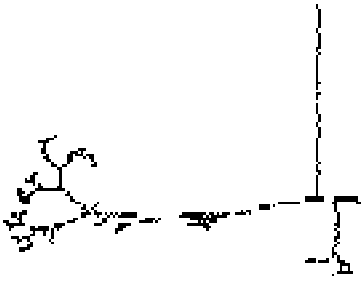


Figure 4
Result of reclassification
of LANDSAT data
(61,440 observations)

Characterization on the Parameter Sub-Family of Hierarchical Clustering Algorithms and the Space-distorting Properties

Nagatomo Nakamura

The Postgraduate Course of Science and Technology
Nihon University
Tokyo Japan

Noboru Ohsumi

Department of Statistical Methodology
The Institute of Statistical Mathematics
Tokyo Japan

The objective of this paper is to present algebraically the agglomerative hierarchical clustering algorithms (the AHC algorithms) and some properties of the clustering strategies. Generally, the AHC algorithms, in which the dissimilarity or distance between a newly-merged cluster $C_i \;\; C_j$ and the remaining

$$d(C_i \;\; C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma \text{½} d(C_i, C_k) - d(C_j, C_k) \text{½} \qquad (1)$$

In this formula, h(C_i) is the height (or level) of cluster C_i a tree structure, such as a dendrogram, and $\theta \equiv \{\alpha_i. \ \alpha_i, \ \beta, \ \gamma\}$ is a set of parameters that characterize the clustering strategies. Some clustering strategies included in the AHC algorithms have the property called "reversals in the tree structure," i.e., it can contain clusters C_1 and C_2 for which $C_2 \;\; C_1$ but produce h(C_1) < h(C_2). With respect to this matter, Lance and Williams (1967) introduced the concept of space distortion, explaining the clustering strategies as space-contracting, space-conserving or space-dilating, but they discussed these concepts intuitively.

On the other hand, Milligan (1979) and Batagelj (1981) present necessary and sufficient conditions for suppressing the presence of reversals produced by the AHC algorithms. Moreover, Dubien and Warde (1979) derive under some assumptions the more sophisticated results of space distortion among distances that are obtained at different levels in merging clusters. However, these efforts are restricted to characterization of a sub-family of AHC algorithms, the (β-γ) family in the algorithm described as expression (1).

This paper, therefore, discusses extension and sophistication of the concept of the space distortion. First, we define the concept as follows:

Letting

$$\Delta = \{d(C_i \;\; C_j, C_k) \ at \ (\alpha_i, \alpha_j, \beta, \gamma) \text{½} d(C_i, C_j) < d(C_i, C_k) < d(C_j, C_k) \qquad (2)$$

be a set of all distances obtained from the result of an agglomeration or a clustering process, we define the conditions of space distortion as:

1) space-conserving:

$$d(C_j,C_k) < d(C_i \quad C_j,C_k) < d(C_j,C_k)\ ;\ d(C_i \quad C_j,C_k) \quad \Delta$$

2) space-contracting:

$$d(C_i,C_k) \geq d(C_i \quad C_j,C_k)\ ;\ d(C_i \quad C_j,C_k) \quad \Delta$$

3) space-dialating"

$$d(C_j,C_k) \leq d(C_i \quad C_j,C_k)\ ;\ d(C_i \quad C_j,C_k) \quad \Delta$$

The relationship between the parameter set θ and the space distortion is clearly examined according to these definitions. Especially, the {α_i, α_j, β, γ} family of AHC algorithms and the regions of the Cartesian plane investigating a strategy for the {α_i, α_j, β, γ} algorithms are strictly derived. Hence, this investigation clarifies the classification of most clustering strategies that are characterized by the general agglomerative algorithms (1) belonging to the {α_i, α_j, β, γ} plane (for example, single-linkage, complete-linkage, centroid method, group average method, median method, Ward's method, minimum variance method by Jambu, and so on). Experimental examples obtained using artificial data sets are discussed. In addition, the relationships between each parameter and the agglomerative clustering strategies are graphically presented within a domain specified by the {α_i, α_j, β, γ} plane.

References

(1) V. Batagelj (1981):Note on ultrametric hierarchical clustering algorithms, Psychometrika, **46**, 351-352.

(2) J. L. DuBien and W. D. Warde (1979): A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms, The Canadian Journal of Statistics, 7, 1, 29-38.

(3) G. N. Lance and W. T. Williams (1967): A general theory of classificatory sorting strategies, I.Hierarchical systems, The Computer Journal, **9**, 373-380.

(4) G. W. Milligan (1979):Ultrametric hierarchical clustering algorithms, Psychometrika, **44**, 343-346.

A New Criterion for Classification Using AID

Toshiro Haga and Chikuma Hamada

Department of Management Science
Science University of Tokyo

When predictor variables which are two-way classification types are used with the AID (Automatic Interaction Detector) program, assigning a score of 0 and 1, or 1 and 2, to the predictor variables makes it possible to use maximization of the sum of squares between the groups or of the variance ratio as the partitioning rule, as when the predictor variables are quantitative variables. Use of χ^2 in a contingency table and of entropy have both been suggested, but the characteristics of these two methods differ little from those of the sum of squares between groups method. This report introduces a new classification criterion which differs significantly from conventionally used criteria.

The validity of a classification result can be judged from the following two conditions:

1) The difference between the average scores (ratios) for groups 1 and 2 is large.
2) The groups are partitioned so that the number of cases in both groups is approximately the same.

Table 1 Cross totals (number of cases)

y	Group 1	Group 2	Total
0	a	b	r_1
1	d	e	r_2
Total	c_1	c_2	n
Mean	p_1	p_2	p

Table 2 Example

Previous	Illness	White Cell	Count	Total	
No	Yes	<7000	≥7000		
86	1	19	68	87	
122	25	70	77	147	
208	26	89	145	234	
p	0.587	0.962	0.787	0.531	0.650
$\lvert p_1 - p_2 \rvert$	0.375	0.256			
	F_0	13.9	15.4		
	F_1	23.2	16.6		
	F_2	54.3	18.1		

When the distribution of cases following classification is as shown in Table 1, the variance ratio between groups is expressed by the expression below. In order to distinguish this variance ratio from the one we will introduce below, we shall call it F_0.

$$F_0 = (p_1 - p_2)^2/\{p(1 - p)(1/r_1 + 1/r_2)\}$$

Under the null hypothesis that the expected values of p_1 and p_2 are equal, the variance ratio is obtained using the variance within the group as that for the entire sample.

Since in partitioning using AID, the classification is conducted under the assumption that the expected values of p_1 and p_2 will be different, a new variance ratio F_2 is defined using the internal variance within each group.

$$F_2 = (p_1 - p_2)^2/\{p_1(1 - p_1)/r_1 + p_2(1 - p_2)/r_2\}$$

Finally, the variance ratio obtained from variance equalization through arc–sine transformation performed on p_1 and p_2 is defined to be F_1.

$$F_1 = (\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{p_2})^2/\{(1/4)(1/r_1 + 1/r_2)\}$$

Comparing the characteristics of the three variance ratios described here, we see that F_0 emphasizes condition 2 above and F_2 emphasizes condition 1. F_1, on the other hand, shows characteristics intermediate between the two. Table 2 shows a concrete example of this.

Using the classification criteria presented here, we constructed diagnosis charts for cerebral hemorrhage patients and cerebral infarction patients from clinical data and obtained specific partitioning patterns for each partitioning criteria. For each partitioning, F_2 tended to produce groups with highly skewed ratios (meaning that the diagnosis can be confirmed) despite the small number of samples contained in a group. F_0 was able to produce groups with skewed ratios only after repeated partitioning. These features were obtained using mathematical simulations.

A Conversational AID System

Toshiro Haga and Motohiro Setoya

Department of Management Science
Science University of Tokyo

It is extremely important that full technical study proper to the field of application be conducted on any set of analysis results. In fact, it might be better said that such a study should be conducted not on the "results" of an analysis, but rather on the "stages" of that analysis. One of the authors of this report has developed or been involved in the development of two programs: CDA (Conversational Data Analysis) and QCAS/MA1 (Quality Control Assisting System/Multivariate Analysis Part 1). Both of these programs are used in analyzing data conversationally on a personal computer. As its name indicates, AID (Automatic Interaction Detector) is a program which conducts an analysis automatically, but by adding his ideas during an analysis, an analyst can obtain binary trees with high applicability to a particular site. In this report, the authors will introduce CID (Conversational Interaction Detector), which operates as a CDAsubsystem.

The stages of conversational processing are as follows:

1. The variables used in classifying and the partition locations are displayed on the screen in order of F ratio. A graph is displayed on the upper part of the screen and the number of cases, mean, classification F ratio, etc., for each group are displayed below that.
2. With the information displayed on the screen and the analyst's technical knowledge, the analyst selects a classification and instructs the computer as follows. To obtain the classification with the largest F ratio, simply enter D (for divide). Thereafter specify D followed by the variable number and classification location.
3. When classification has progressed somewhat, to see the classification up to that point, simply enter T (for tree) and the binary tree will be displayed. Except for branches which have been completely classified, it is possible to experiment with a different classifications. This feature can be used in the following situation: when different (interactive) variables are used for classifying two branches and the analyst wishes to cancel one of the classifications and use the same variable for both.
4. As Haga and Hamada have reported separately, changing the classification criterion will produce different results. Not only is it possible to specify the criterion to be used at the beginning of the analysis, but the criterion can be changed during the analysis in order to obtain the optimum classification.
5. As with the CART (Classification And Regression Tree) program, it is possible to specify whether the category order will maintained for each variable during classification and tospecify multiple classification stopping rules.

An example illustrating the merits of the conversational method follows. In diagnosing patients with cerebral hemorrhage, the variable "Does your head hurt? often includes missing value and cannot be used in classification. This is because many patients are unconscious, and thus, it is not possible to obtain an answer. However, the variable "Does your head hurt?"can be used after a classification using the variable "Is the patient conscious?". In conversational processing, it possible to give full consideration to this type of problem while progressing through the classification.