Bulletin
of
The Japanese Classification Society

Volume 2

December 1988
JAPAN

Contents

Recent Developments in Quantitative Analysis of Japanese Documents

Masakatsu Murakami
The Institute of Statistical Mathematics
4-6-7, Minami-Azabu, Minato-ku
Tokyo Japan

Since the development of computer technology has enabled us to analyze Japanese characters such as Kanji and Kana, the quantitative analysis of Japanese documents has come to a new stage.

Some data bases of Japanese literature are already constructred in the Data Processing Center of Kyoto University and Institute of Investigations on Japanese Literary Documents; however, the analytical methods for using these data bases have not been developed.

Since 1974, we have constructed a data base for Nichiren's documents to solve the questions of authorship of five disputed documents. Nichiren (1222~1282) opend a new school of buddism called Nichiren-shu in the Kamakura era.

The purpose of this paper is to introduce a method of quantitative analysis using Nichiren's papers and to point out the difficulties of the quantitative analysis of Japanese documents.

The papers we use for the quantitative analysis are as follows.

Nichiren's papers,	24	(85000 words)
Forged papers,	16	(15000 words)
Papers of Nichiren's students,	5	(13000 words)
Disputed papers,	5	(20000 words)

Using these papers, we try to distinguish the characteristic feratures of Nichiren's sentences. We focus on the following two different types of information for each paper.

1) The frequency of all the words.
2) The information about a literary style such as the distribution of sentence length, the distribution of word length, the mean of sentence length, the mean of word length, the type-token ratio, the ratio of parts of speech, Yule's K-statistics, etc.

In addition to our quantitative analysis of Nichiren's papers, we also discuss the difficulties of analysis of Japanese documents.

Analysis of the Growth Data According to Age on Boys and Girls - An Aspect of the Growth Pattern of Height -

Reiko Ninomiya
Computational Laboratory
Japan Women's University

Yukiko Higuchi
Shikiko Hayakawa
Faculty of Home Economics
Japan Women's University
1-2-8, Mejirodai, Bunkyo-ku
Tokyo Japan

As a result of our research on body form classification of boys and girls and it's variation in growth of adolecent, we grasped variation of body form and individual difference of growing-up speed were remarkable.

In this report to analyze individual difference of growth, we used the data of personal growth by age.

We classified growth patterns. Concerning the age of the maximum growth, we discussed classified patterns, growing-up amount in adolecence term, height of first grader in elementary school and growing-up amount of the age of the maximum growth.

The results are as follows:

1) The classified patterns were characterized by the age of the maximum growth, level of growing-up speed and difference of growing-up speed by age.

2) There were significant differences between growing-up amount of the age of the maximum growth.

3) The growing-up amount in adolecent term were significantly greater for boys and girls whose age of the maximum growth were later than for boys and girls whose age of the maximum growth were earlier.

4) The height of first grader in elementary school were significantly smaller for boys whose age of the maximum growth were later than for boys whose age of the maximum growth were earlier.

Presence of Subtypes of Anxiety Neurosis as Revealed by Classifying its Symptoms by the Mini-numerical Taxonomy System

Yuko Suzuki, Koujirou Shukutani, Shizuo Aizawa, Shinobu Tatsunami*, and Nagasumi Yago*

Department of Neuropsychiatory and *Radioisotope Research Institute, St. Marianna University School of Medicine, Sugao 2-16-1, Miyamae-ku, Kawasaki-shi 203 Japan

Number of serious cases of anxiety neurosis have increased in recent years in Japan. Symptoms shown by these patients may not be improved either by drug administration or by phychiatric therapy. For the purpose of finding out possible factors that underline the symptoms, we have constructed and analyzed a data file on personalities of 55 patients of anxiety neurosis including 14 patients of the serious cases. The raw data which were collected by means of the University of Tokyo Personality Inventory (abbreviated as TPI) were first transformed into personality scores according to the handling manual of TPI. Then, the file was analyzed by the Mini-Numerical Taxonomy System (abbreviated as MINTS) which has been developed by Dr. Ohsumi at the National Institute of Statistical Mathematics, Tokyo. Using the group average method, we could classify the patients on the dendrogram into 10 clusters which were numbered from 1 through 10. Of the 10 clusters, Clusters 1, 2 and 3 clearly composed the major group while Clusters 4, 5 and 7 the other. A characteristic feature of the major group was that the patients showed good prognosis. Patients that belonged to the second group did not show any improvements by pscychiatric treatments. These patients were found either hysterical or epileptical. In conclusion, the cluster analysis of symptoms by the MINTS was found to be quite useful in diagnozing patients of anxiety neurosis especially in preliminary classification of the patients into several subtypes.

Patient Classification for Medical Resource Allocation

Takao Ishizuka
Faculty of Business Management
Asia University
24-10 Sakai, 5-chome, Musashino-shi
Tokyo Japan

Medical resources should be allocated according to needs of patients. Amethodtoclassify patients into some groups on the standpoint of resource allocation is called “casemix classification.” Patient classification by the case mix studymightbeveryimportantforthe hospitals from now on

The objective of this paper is to summarize several patient classification systems including DRG which is adopted for Medicare Prospective Payment System in U.S.A., and propose a graphic case mix analysis method by posterior probability of some factor affecting the variation of length of stay

It is demonstrated that this method is quite useful as:

1) quantization of the extent of separation of mixed distribution,
2) visualization of case mix complexity, and
3) optimal selection of variables for classifying cases into homogeneous groups from the statistical and clinical view.

Evaluation of River Water Quality of a Mountains Ecosystem by Cluster Analysis

Tomoyuki Hakamata, Tatemasa Hirata and Kohji Muraoka

Water and Soil Environment Division
The National Institute for Environmental Studies
Tsukuba-shi, Ibaraki 305
Japan

River water is sink of cycling materials in an area ecosystem, source of which is precipitation. Because the river water quality reflects a lot of biogeochemical phenomena, the land use effects in the ecosystem may be assessed by exploring the river water quality. In this study, cluster analysis was used to evaluate the river water quality in a mountains ecosystem which have many typical land-use, e.g., forest, paddy field, upland field, orchard, village, etc.

The 153 sampling sites of the rivers were selected in the Tsukuba mountains ecosystem. The sample water was collected in late winter or spring of 1986. The land use types of the sites were recorded at the same time.

A frequency distribution of silica(SiO_2) was symmetrical. However, that of pH, elctric conductance(EC) and other elements was asymmetrical, the smaller class having much higher frequency. The medians of the data were compared with those of Kanto district reported by Kobayashi(1960). The medians of nitrate nitrogen(NO_3) and chlorine obtained from this ecosystem were higher than or nearly equal to those of Kanto district. Nitrate nitrogen is assumed to be increased by biological processes in forests and agricultural land of the ecosystem. Chlorine is assumed to be moved into the ecosystem from the Pacific Ocean placed in 40km eastern part of Mt. Tsukuba.

The sampling sites were classified into 10 groups by the Ward method of cluster analysis from the viewpoint of water quality and land-use. The groups is reflected degree or pattern of human impacts (Table 1). Phosphate(PO_4), nitrite nitrogen(NO_2) and NO_3 are concerned with human life. High EC value is often observed in water near which menactivelymove.SiO_2 is connected with quarrying characteristic in this region.

Table 1. The classification of sampling sites from the viewpoint of water quality and land-use

Groups	Numbers of samples	Characteristics of water quality	Location or land-use
(1)	12	Low pH, lowest SiO_2 & middle NO_3	Near mountain top & forest
(2)	11	Lower EC/NO_3 & higher SiO_2 than (1)	Upper mountain & fore
(3)	11	Higher PO_4 SiO_2 than (1) & (2)	Lower mountain & mainly forest
(4)	12	Higher EC & lower PO_4/SiO_2 than (3) & near road	Upper outside of village
(5)	21	Lower $EC/PO_4/NO_2$ than (3) & (4), middle $pH/SiO_2/NO_3$	Mountain & forest
(6)	12	Lower pH & higher EC/NO_2	Near village or near main road
(7)	13	High NO_2/NO_3	Lower outsides of village
(8)	10	High NO_2/NO_3 & low pH/SiO_2	Upper parts of main river
(9)	7	Highest SiO_2, high pH/EC & low NO_2/NO_3	Village or under quarry
(10)	2	High SiO_2, low or middle others	Lower outsides of large village or lower parts of main river

Some typical cases in which natural riverbeds concserve water quality were observed: for example, in R. Kabaho three samples from a branch with artificial riverbeds (ground sill) were classified in the group (5), whereas three samples from another branch with natural riverbeds were classified in the group (1). Natural riverbed should be protected even when forests around the river are used for any purposes.

Toward a Genuine Cladistics

Kuniyoshi Ohta

Seiroku Sakai
Daito Bunka University
Higashi-Matsuyama, 355
Saitama-ken, Japan

Systematic biology is one of the oldest among the sciences of classification (e.g. Knight, 1981; Hachiuma, 1987). This paper, based upon Ohta's (1988) study on the exact history of systematics theory, gives a brief account of the main current schools of systematic biology. It is pointed out that Smirnov's (1968) “exact systematics” should be clearly discriminated from Sokal and Sneath's(1963) “numerical taxonomy”, because the former is not concerned with overall similarity.

It has lately become the fashion, at least among the German- and English-speaking systematists, to advocate the Hennigian cladism (Hennig, 1966). The second part of this paper criticizes cladism or Hennigism in the following points (Ohta, 1988)

(1) The Hennigian or cladist regards W. Hennig as the founder of phylogenetic systematics and of cladistics. This is, however, completely wrong. The founders of phylogenetic systematics are Darwin, Haeckel, and others, and the founder of cladistics is P. C. Mitchell (1901, 1905).
(2) The Hennigian regards systematics, phylogenetics, and cladism as all equivalent, but this is incorrect.
(3) Hennigism or cladism is not true cladistics.
(4) “Hennig's principle” of identification of sister relationships by means of synapomorphy is inaccurate in words and logic. It should be replaced by what we call Mitchell's Principle of identification of most relatives by means of shared, uniquely derived characters.
(5) Hennig's classification and definition of ”monophyly”, “paraphyly”, and “polyphyly” are fundamentally errneous. There is not a third alternative to monophyly and polyphyly in the true sense--logical and historical--of the terms defined by Haeckel.
(6) A monophyletic group in the true sense is a phyletic group such that a global root species (G) is included in it.
(7) Hennigian's “monophyletic group” seems to be our hologroup, but Hennigian's thesis that only their “monophyletic groups” are taxa is a narrow approach. We recognize semiphyletic as well as holophyletic groups as taxa.
(8) Hennigian's ideology of dismissing ancestor-descendant relationship is wrong.

The third part of the present paper, based on Ohta's(1988) work, gives a set of rigorous definitions and measures of four components of phylogenetic relationship which has been so far very vague in systematic biology.

(i) Phylogenetic relationship is composed of generalogical and genetic relationships, and the former composed of cladistic and chronistic relationships.

(ii) The cladistic relationship in the strict sense between any two species in a phylogenetic tree is defined by the triplet (k, l;m), where k and l are respectively the species-generation numbers between each of the two species and their local root species (L), and m is the species-generation number between L and G. Evidently m is a measure of predivergence kinship between the two species, while the vector (k,l), or the scaler $(k+l)$, might be useful for representing a measure of postdivergence kinship between the two species. For example, two sister species in the true sense corresponds to the relationship in the (1,1)th degree.

(iii) The chronistic relationship between two species is defined as the triplet (*r,s;t*), where *r* and *s* are respectively the year numbers between each of the two species and L, and t is the year number between L and G.

(iv) Genetic relationship is defined as the degree of genotypic homology between two species.

The last part of this paper offers a probablistic illustration of Mitchell's Principle for identification of closest relatives. Consider a simplest, irreversable transformation series, A_0, A_1, where A_0 and A_1 are respectively the primitive and derived characters, and the transformation probability is assumed to be α in unit time. Then the probability that two species in the relationship of (*r,s;t*) have in common the derived character A_1 is

$$P(A_1|r, s; t) = 1-(1-\alpha)^t+(1-\alpha)^t[1-(1-\alpha)^r][1-(1-\alpha)^s],$$

where the first and second terms represent the probability of homology, and the last term represents the probability of homoplasy. On the other hand, the probability that the same two species have in common the primitive character A_0 is

$$P(A_1|r, s; t) = (1-\alpha)^{r+s+t}$$

Here, we define the information contents of shared, derived characters and of shared, primitive characters respectively as

$$I(A_1|r, s; t) = -\log P(A_1|r, s; t),$$

$$I(A_0|r, s; t) = -\log P(A_0|r, s; t).$$

Then it will be easily shown that the information content of shared, derived characters will increase as $\alpha \to 0$, and vice versa for shared primitive characters. This is an information-theoretic foundation of Mitchell's Principle.

Furthermore, we shall consider the problem of estimating the genealogical relationship by using shared, uniquely derived characters. For the purpose, let $P(t \mid A_1)$ be the probability that the degree of relatedness of two species sharing A_1 is t, and $R(t)$ be the probability of *t* in any two species within a phylogenetic tree. Then we have

$$P(t \mid A_1) = \frac{P(A_1|t)\, R(t)}{\sum_w P(A_1|w)\, R(w)}$$

and if we can assume that $R(t)$ is a uniform distribution and that A_1 is a uniquely derived character, we obtain a very simple theorem.

$$P(A_1|r, s; t) = 1-(1-\alpha)^t+(1-\alpha)^t[1-(1-\alpha)^r][1-(1-\alpha)^s],$$

This is a probablistic foundation of Mitchell's Principle. The final part of the present paper points out that there is a strong need for an Integrated Taxonomy (Sakai, 1980) or Synthetic Systematics (Ohta, 1988) to resolve the endless controversy among the current systematic schools and to advance a new framework in theoretical systematics.

References

(1) Hachiuma, T. (1987) The Dawn of Theoretical Classification Science (In Japanese), Takeda Shoten, Fujisawa.
(2) Hennig, W. (1966) Phylogenetic Systematics, Univ. of Illinois Press, Urbana.
(3) Knight, D. (1981) Ordering the World, Burnett Books, London.
(4) Mitchell, P.C. (1901) Trans. Linn. Soc. Lond. Zool., 8: 173-275.
(5) Mitchell, P.C. (1905) Trans. Zool. Soc. Lond., 17: 437-536.
(6) Ohta, K. (1988) In: Zool. Soc. Jap. (ed.) Evolution--the New Synthesis, Ch. II (in Japanese), Gakkai Shuppan Center, Tokyo.
(7) Sakai, S. (1980) Insect Integrated Taxonomy, Daito Bunka Univ. Takasaka.
(8) Smirnov, E.S. (1968) Taksonomicheskii analiz, Izd. Moskov. Univ. Moskva.
(9) Sokal, R.R. and P.H.A. Sneath (1963) Principles of Numerical Taxonomy, Freeman, San Francisco.

A Visual Data Analysis System

Masahiro Mizuta
Division of Information Engineering
Hokkaido University
North 13, West 8, Kita-ku, Sapporo-shi
Hokkaido-060, Japan

The efficient manipulations of data analysis system have begun to arouse considerable attention with the popularization of the systems. The menu method and the command method are ordinary manipulations of them. Another methods are developed with the advance of the study of man-machine interface. One of the most attractive method is the utilization of graphics, in other words vision. Window system and ICON of work stations or Macintosh are typical examples of them. Visual programming environments have been studied in the field of the information engineering.

The conception of the visual programming is applied to the field of the data analysis and we get a view of visual data analysis. Some of the methods of data analysis use graphics, which are called graphical methods. The objective of the paper is to visualize the manipulations of data analysis. We produced a visual data analysis system on a micro by way of trial. An outline of the system is mentioned in the following.

Any procedure of data analysis, for example, read from a file, computation of variances or PCA etc., is considered as an ICON abstractly. The ICON have some INPUTs and some OUTPUTs. It has also some OPTIONs i.e. parameters, which specify the action of the procedure. In the case of cluster analysis, the ICON has an INPUT, an OUTPUT and an OPTION. The INPUT implies similarity matrix, the OUTPUT is a set of labels of objects, the OPTION specifies the method of cluster analysis and parameters. The sequences of data analysis can be expressed the sequences of the ICONs. The trial product of a visual data analysis system is depended on this iconic model.

Anormal manipulation of the system will be explained here.

(1) Click the place of 'ICONS' with mouse. The list of ICONs which is registered, appear immediately. Select an ICON with mouse.
(2) Place the ICON on the Work Area on the screen. Repeat step 1 and 2.
(3) Connect an OUTPUT of an ICON to an INPUT of another ICON by lines.
(4) If necessary, set OPTIONs of ICONs.
(5) Execute each ICON in order, according to click 'EXEC' on the ICON with mouse.
(6) Change the ICONs or the connection of the ICONs on the screen, for trial and error.
(7) Terminate the system by QUIT.

One of the advantages of the system is flexibility of the procedures correspond to ICON. The registration of the procedures is easy, because each procedure is independent program. The body of the system communicates with the procedures by two ways. The first is the utilization of the specific file and the second is of the argument list. The program of a procedure can get the informations from the body of the system with only one of these ways.

A construction of a data analysis system with plain mainpulation is not plain. However, the use of visual is very hopeful for easy operation of the system. We report one system which is produced with the intention of furnishing a field of programs.

Application of the Graphic Software GDAS for Classification Problems

Kiyoshi Katayama, Tugihiro Shu
Keiji Yajima
Institute of JUSE
4-30-3, Sendagaya, Shibuya-ku
Tokyo, Japan

Newly developed software product GDAS(Graphical Data Analysis System) aims to carry out data analysis by means of graphical presentation on the microcomputer. There have been a couple of graph drawing program products and the GDAS is characterized by strong orientation toward data analysis through fair drawings.

To illustrate the classification capability some branches such as so-called radar chart, diagonal presentation of three variables(%), and the time series plotting are shortly introduced.

For example, in radar chart utilizing a scale adjustment facility a comparative study of the samples may be provided.Feature of computer display is characterized by easily tackled trialsof a number drawings, with the size alteration and the sample batch changes.

In the time when there is an overflowing data supply, one ought to put the software in full use of the classification techniques.