Invited Paper 28.3
46th Session of the ISI

STATISTICAL INTELLIGENT SOFTWARE FOR AUTOMATIC CLASSIFICATION

Keiji Yajima
Institute of Japanese Union Scientists and Engineers
Tokyo, Japan

Noboru Ohsumi
The Institute of Statistical Mathematics
Tokyo, Japan

1. Introduction

A quarter of a century has already elapsed since the development of statistical computing systems and statistical software made a start in the 1960s. At present, we have easy access to a variety of statistical systems such as the BMDP, GENSTAT, SAS, and SPSS-X, and can also find that the computer environment in which these systems are used has been sufficiently improved in all aspects. We can say that all these systems are based mainly on the principle of batch or semi-batch processing system which means that the information flows in one direction only. Recent years have also seen a series of heated debates among statisticians and data analysts over statistical intelligent system and expert system. (Gale, 1986; Haux, 1986; Hand, 1984)

In this report, the problems in developing intelligent statistical software or expert system for automatic classification are discussed to find a clue to obtaining the prospects for automatic classification in the 1990s. A prototype of such an intelligent system and its design concept are also proposed, and an experimentally developed hardware configuration is introduced.

The term "Statistical Expert System" may be defined in two different ways, depending on the functions it is expected to perform. In one way, the system is expected to operate as a support tool for accelerating the sophistication and evolution of statistical software. In the other, it is expected to function as a problem-solving tool, or as a true expert system, that can play the role currently assigned to statisticians.

No consensus of expert opinion has yet been reach as to what exactly the statistical expert system means and what functions it should have. It can be safely said, however, that any attempt to find a remedy for the drawbacks in the existing statistical systems will invariably lead to creative development efforts focused on the building of an innovative statistical expert system.

While this holds true with automatic classification software, we know that automatic classification involves more unsolved problems than does statistics, some of which are enumerated below.

1) There are innumerable techniques of automatic classification, and the concepts adopted for developing such techniques are derived from a great variety of theoretical fields (e.g., statistics, graph theory, combinatorial theory, mathematical programming, fuzzy theory).

2) The research areas to be covered and utilized are very extensive, and the application areas are consequently made very widespread.

3) For these reasons, confusion and jargon in terms, techniques and algorithms have arisen.

4) Many of the techniques and theories are rather heuristic and exploratory ones, and are not always given satisfactory mathematical explanations.

5) Despite the widespread use of such techniques and theories, there are few examples of their successful application. In other words, there is a lack of accumulation of empirical knowledge of high quality level.

It is probable that these problems will hinder the categorizing of the problems in automatic classification to be solved by developing intelligent software, the determination (selection) of a group of specialists, and the systematization of expert knowledge, all of which are essential and fundamental procedures necessary for building an advanced knowledge base. Although it is certainly difficult to attain the construction of implementing an advanced expert system in the immediate future, we can nevertheless foresee the possibility of implementing it by stages. For the purpose of such step-wise approach, it will be necessary to pursue the following steps.

1) Upgrading of existing automatic classification software to intelligent software: The knowledge base and individual parts of the inference engine will be generated by making use of existing software functions such as the dedicated command language, meta-languages, macro commands, etc. to realize self-growing development of the existing software system to an intelligent system.

2) Software sophistication using front-end processors: The existing software functions (e.g., techniques, database, data modification, editing) will be used either for execution or as parts of the inference engine, and in addition, an intelligent front-end processor system will be built which is capable of allocating the execution process of a user-demanded job task. This will be linked with the existing statistical system so that the front-end processor will convert each command to an executable form by translating it to a form readable by the statistical system according to the user's demand. This processor portion will be registered in the knowledge base as far as practicable.

3) Intelligent system creation using tools for building expert systems: This step is an approach advanced from step 2) above. Specifically, the knowledge base and the inference engine will be designed using a suitable support tool for development (i.e., expert shell) by capitalizing on existing software, particularly the accumulation of know-how in highly integrated systems (e.g., CLUSTAN, NTSYS, SICLA, MINTS).

4) An expert system based on an entirely new concept will be developed.

In the following, a prototype of intelligent software for automatic classification is conceptually introduced. It has been designed with consideration given to the above development steps as well as to the existing state of automatic classification.

2. Present Situation of Users and Experts in Automatic Classification

As pointed out already, one of the outstanding features of automatic classification is that it is utilized by researchers in many different fields and there are a large number of experts specializing in its research. This means that users (or experts) will show their understanding for automatic classification in varying degrees and within varying ranges. An expert specializing in classification may use a computer, though he is not skilled at its operation. A researcher having no knowledge at all about classification techniques may engage in data analysis using a computer for his specialized research area or subject. Furthermore, a user who is an expert in certain field may be an absolute beginner in certain other fields, and he may not be computer-experienced either.

Considering the difficulty in building a system which can meet users' demands of all types and is at the same time compatible with an environmental conditions under which it is used, it may be advisable to classify users into three groups, i.e. beginners, middle-level users and experts, according to the level of their knowledge about automatic classification. This will enable each individual user to input his own level to the computer according to his own judgement, thus providing him with the freedom of using the system in the most appropriate manner. Such consideration in the design of the user interface will be required because the definition of "expert" in automatic classification still remains ambiguous. By reason of the factors mentioned above, it is likely that in the course of the system construction, difficulties will arise which differ from those encountered in the development of other systems that have relatively clear-cut objectives, such as traffic analysis expert systems and medical diagnostic expert systems, and these difficulties are essentially ascribable to the rapid pace of development of diagnostic knowledge and related information.

3. Knowledge Representation in Automatic Classification Software

It is the accumulation of expertise in the form of rules that constitutes the basis of knowledge representation. As stated already, however, no such environment has yet been created that is sufficiently upgraded for systematization of the expert knowledge of classification or the contents of classification research. What can be done at the moment would therefore be limited to the classification and summarization of routine analytical procedures (or empirical rules) according to the degree of their clarity. For example, the rules may be considered in a number of categories, as explained below.

(1) Rules that are mathematically or theoretically substantiated

The relationship among distances used in the combinatorial hierarchical classification is an example of such rules. In the case of Ward's method, for example, the squared Euclidean distance alone is usable. In the case of complete-linkage and single-linkage methods, on the other hand, it is allowable to use rather general dissimilarity (or similarity) coefficients so long as the symmetry property is satisfied. As another example, furthermore, the k-means method is applicable only to quantitative data, and requires specifying criteria of homogeneity.

(2) Rules that can be acknowledged as facts or characteristics, but not clearly substantiated by mathematical explanation

It is known that many of hierarchical methods produce results that show a monotonic hierarchical structure and ultrametric properties. Some of such methods cause the reversal of linked distance, but the relationship between such properties and the given data structure has not been made clear yet.

(3) Rules that are mathematically unclear or heuristic, but are frequently used

Rules of this kind include numerous formulas of similarity (or dissimilarity), classification methods by asymmetric similarity matrix, and evaluation criteria of hierarchical structures derived from hierarchical classification methods.

The design policy of the knowledge base must be formulated by giving careful consideration to the rules cited above. Specifically, a clear distinction should be made between the cases where the production rules (If premise/condition Then conclusion/action rule) are suitable and the cases where the frame-based methods are suitable.

For description of general characteristics of classification techniques, knowledge representation based on the production rule is suitable. The characteristics referred to here indicate the items listed below, which have hitherto been given as optional functions of automatic classification software, but should be stipulated more clearly as rules on the basis of their compatibility and relationship among data and/or methods.

- Main application areas
- Data attribute (quantitative/qualitative, nominal/ordinal/proportional/interval, transformed data/aggregated data)
- Data table type (cases by variables, cases by cases, etc.)
- Size of data (large, medium, small)
- Methods used
  - Primary stage (hierarchical, non-hierarchical)
  - Secondary stage (agglomerative type/divisive type of hierarchical method, partitioning type of non-hierarchical method)
  - Others (combinatorial type, hybrid type, fuzzy clustering, etc.)
- Selection of similarities/dissimilarities
- Selection of optimization criteria
- Selection of algorithms
- Evaluation, interpretation, diagnosis of classification results (statistical evaluation criteria, number of clusters, evaluation among dendrograms, etc.)
- Output representation method (graphical type, tabulated type, summarization, etc.)

For example, the knowledge representation of the following rules can be conceived.

Exhibit I.

IF
- Type of method = (hierarchical) and (agglomerative)
- Data attribute = (qualitative data) and (multicategory data)
- Data table type = (multivariate data (cases by variables))
- Size of data = (small) or (medium),

THEN
- Methods or kind of processing = (combinatorial hierarchical classification)
  or
  (graph-theoretic method(s))
  Example: Single-linkage method
  Complete-linkage method
  Minimum spanning tree generation
  Ling's (k, r) method
  .....
  .....

and
- Similarities = Applicable to nearly all kinds of similarity
- Output representation = Graphical representation (e.g., dendrogram) is suitable, and availability of clustering summarized table (membership list, statistics, goodness of fit) is desirable.

and
- Interpretation = (Attention should be directed to the following points)
  - Cluster form or data structure should not be judged from the dendrogram's form alone.
  - Relationship between the similarity used and the classification algorithm cannot be made very clear.
  - Compatibility between the dendrogram and the similarity should be carefully evaluated, if necessary.
  - Suitable transformation/modification of input data may be required (e.g., binarization processing).

Exhibit II.

IF
- Object of classification = (remotely sensed data)
- Type of method = (partitioning type)
- Data attribute = (quantitative data)
- Size of data = (extremely large) or (medium)

THEN
- Methods or kind of processing = (The following can be cited as alternatives)
  - k-means type procedures
  - ISODATA method
  - Fuzzy partitioning techniques
  - Dynamic clustering approach

and

Similarity = (Strictly limited to squared Euclidean distance)

and

Criteria of optimization = (Designation is required)
- Sum of squares criterion (trace criterion)
- Determinant criterion
- Total sum of squared Euclidean distance within clusters

and

Evaluation, interpretation = (Attention should be directed to the following points)
- Initial partition (random start, systematic assignment, use of part of given data, use of seed points, etc.)
- Up- and downdating formulae, relocation formulae
- Suitability to detection of well-separated clusters

4. Utilization of Frame-Based Knowledge Base

All of information related to heuristic findings and mathematical evidence obtained from the characteristics or trends of each classification method or from the past experiments and experience can be registered as "facts" in the frame-based knowledge base system. The description of phenomena incidental to the analytical process of clustering or the empirical knowledge of such phenomena tends to be rather ambiguous and contain just fragments of information. The "frame" is therefore suitable for storage of such information which includes the following items.

Chaining effect
- Objects on dendrogram is linked in chain-like form.
- Chaining effect is occasionally produced when the single-linkage method is used.
- Not many well-separated clusters can be observed, and intercluster contact and bridging phenomena are observed.
- The cluster size is not uniform by grouped.
- Even when the chaining effect is observed, separable clusters caused by other method (e.g., MST - Minimum Spanning Tree) are occasionally observed.

Wild shots
- Search in the neighborhood of cluster centroids discloses that the density of data points is low.
- Distances among clusters are far smaller than those between objects within a cluster.

Influence of outliers
- Abnormally small cluster sizes are observed.
- In particular, a large number of singletons are observed.
- Any change in the methods or initial conditions causes the clustering result to vary largely.

Selection of optimization criteria
- Sum of squares criterion (criterion for minimizing squared deviations within clusters, also called trace criterion).
  . No consideration is given to the correlation between variables.
  . Cluster size does not show an excessive lack of uniformity.

. The criterion bears on the maximization of sum of squared deviations between clusters.

- Determinant criterion
  . Correlation between variables produces a large influence.
  . Dispersion within each cluster and its direction also produce an influence.

The descriptive items listed above very often prove to be ambiguous knowledge lacking in accuracy. Accordingly, it may be advisable to use such knowledge, not as a permanent knowledge base, but as a temporary one. Furthermore, representation of such knowledge in structured form, i.e., by a network or tree structure, cannot be realized with ease. It may therefore be proposed to store such knowledge temporarily in "blackboards" or "file cards" which are linked directly with the knowledge base editor for access to be made according to the need by the inference mechanism. In this case, the user interface will have to be provided with environmental conditions suitable for interactive knowledge registration and retrieval using a microcomputer, etc. In this way, any ambiguous knowledge or current information that cannot be described as a production rule can be linked with the knowledge base through the user interface.

5. Inference Mechanism

It can be generally said that there are two major approaches automatic classification process. One is the top-down approach in which a certain, specific method is applied to the given data set to probe minutely into the characteristics of the data set by changing various conditions established within the limits set by the method used. The other is the bottom-up approach which is based on the idea that clustering process means to generate clusters that can satisfy certain conditions established according to a certain criterion (i.e., clusters are something to be generated). In this approach, a clue for grasping the data structure is obtained from the characteristics of various clusters thus generated. We cannot say which of the two approaches is correct or incorrect because it is considered both natural and reasonable to accept both, one for a forward approach to inference and the other for a backward approach to inference.

If the user is fairly well informed of the nature, attributes, type and size of the given data set as well as the limits set to and conditions for using the method he prefers, and he wishes to know the analyzing process (i.e., solution path) satisfying all such conditions, he would choose the "forward" approach to inference (i.e., top-down approach), provided that there are a number of alternative parts all leading to a successful solution. For example, if biological systematic classification is the main objective and the user wants to obtain various hierarchical genealogical trees, he would explore the forward approach to meet this requirement.

On the other hand, if the user has an image of his own about the clustering condition (a certain hypothesis he has built up regarding the data structure) or about the desirable cluster form and wishes to select an analyzing process that can meet his expectation, he would find that the "backward" approach to inference (i.e., bottom-up approach) is suitable for his purpose. The backward approach will also be chosen if the clustering method or the result representation method is determined in advance and an analyzing process compatible with such method is to be sought.

6. Development Environment of Expert System

If the proposal advanced above is to be implemented in a concrete form as a system, the following conditions will have to be satisfied for both hardware and software.

1) Automatic classification software has already been developed and acquired, and there is a sufficient accumulation of relevant technical knowledge and know-how.

2) Support tools (shells) for developing intelligent software (or expert system) are made available. In particular, expert shells ensuring connectivity and compatibility with the existing integrated software have been discovered and made available (e.g., rule-based system, frame-based system, language, degree of meta-language).

3) Computer environment
The workstation is basically used as a standalone machine and is linked with the mainframe and microcomputers. In particular, the workstation and microcomputer is designed for free use of window management software, bit map display and mouse, and the outline processor or idea processor is used for smooth, efficient interaction with the user interface, particularly with the file cards and knowledge editor.

A system configuration designed with account taken of these requirements is shown in Figure 1 - The ACTIVE Workstation. The term ACTIVE is an abbreviation for "Automatic Classification Techniques for Interpretation, Visualization and Evaluation." As seen in Figure 1, the system is made up of the minicomputer network, microcomputer network and mainframe network, with the workstation as its core. The following can be cited as principal features of this system.

- A variety of languages can be used (FORTRAN, C-language, PASCAL, LISP, etc.).
- Free linkage between each component in the system is ensured.
- Linkage of the workstation with the microcomputer/minicomputer is especially easy.
- Graphics primitives and integrated functions are greatly upgraded.

(For details of functional performance of the ACTIVE Workstation, refer to Reference (5) (Ohsumi, 1987))

7. Conclusion

In the foregoing, the problems in developing intelligent software for automatic classification have been discussed, and a proposal has been advanced which pertains to the fundamental approach to the building of intelligent systems.

We are aware that much more time and labor will have to be spent before an intelligent system based on our design policy is developed and brought into practical operation. While system development efforts have been made consistently over the past years in a trial-and-error method using the ACTIVE Workstation, our creative energies are currently concentrated on the design of a prototype intelligent system which will be developed along the policy described below.

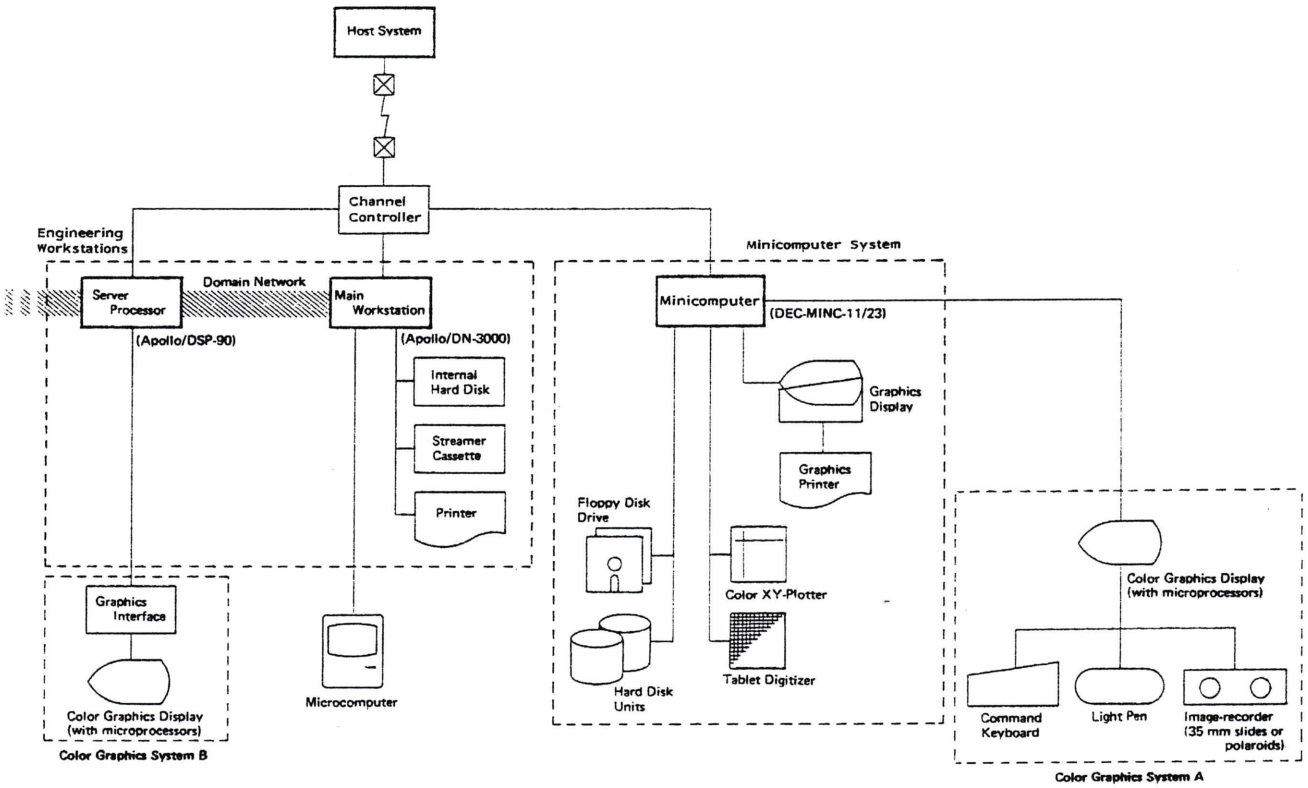

Figure 1 Configuration of ACTIVE Workstation

1) The system will be developed to incorporate a certain limited method (e.g., combinatorial hierarchical method alone) rather than multiple techniques of automatic classification.

2) It is considered inevitable that the selection criteria and contents of the information (i.e., rules and facts) to be registered in the knowledge base will be determined from a subjective point of view.

3) Collection of practical and helpful case studies corresponding to judicial precedents and medical diagnostic data will be started at some later date because it is foreseen that considerable difficulty will be encountered in the course of its implementation.

4) In the development of the knowledge base, therefore, specific importance will be attached to the characteristics of the method to be incorporated, cautions in using the method, compatibility between the method and input data, and guidelines for determining processing conditions.

5) Unlike the case with the integrated software that has already been developed, each function of the method mentioned above will be segmented into small modules to use them as a set of primitives (i.e., individual functions will not be integrated as has been the case with past software development, but will be segmented). These segmented functional modules will be used for automatic generation of analytic procedures needed by the user by making use of the front-end processor linked with the knowledge base.

6) Greater ease of use will be assured for the user interface by making the best possible use of the functions of the workstation and microcomputers as well as the techniques/know-how of window manager, graphics, outline processor and visual programming.

At present, researcher opinions divide on the question of practical application of Artificial Intelligence (AI), some advancing a skeptical view of it, while others express optimistic expectations for its realization. It is quite obvious, however, that the introduction of AI will serve at least to abate much of discontent currently felt with the existing classification software which is designed basically on the principle of batch process system. In other words, a brighter prospect is promised by the future AI application for software performance improvement than by its introduction for general business purposes. From this viewpoint, it can be predicted that the fifth generation computer will make a great contribution to the functional improvement of statistical systems in the 1990s.

BIBLIOGRAPHY

(1) W.A. Gale (ed.) (1986): Artificial Intelligence and Statistics, Addison-Wesley Publishing Co.

(2) D.J. Hand (1984): Statistical Expert System: Design, The Statistician, Vol. 33, 351-369.

(3) R. Haux (ed.) (1986): Expert System in Statistics, Gustav Fischer.

(4) V.D. Hunt (1986): Artificial Intelligence and Expert Systems Sourcebook, Chapman & Hall.

(5) N. Ohsumi (1987): Role of Computer Graphics in Interpretation of Clustering Results in The Proceedings of the Japanese-French Scientific Seminar on "Recent Developments in Clustering and Data Analysis", C. Hayashi and others (eds.).

SUMMARY

In this report, the problems in developing intelligent statistical software or expert system for automatic classification are discussed to find a clue to obtaining the prospects for automatic classification in the 1990s. A prototype of such intelligent system and its outline are also proposed, and an experimentally developed hardware configuration is introduced.

LOGICIEL STATISTIQUE ET INTELLIGENT POUR LA CLASSIFICATION AUTOMATIQUE

RESUME

Dans ce rapport, les problèmes du logiciel statistique et intelligent ou du système expert en développement sont examinés pour trouver la clé susceptible de rendre précises les prospectives de la classification automatique dans les années 1990. Un prototype de pareil système intelligent et sa grande ligne sont aussi proposés et la configuration du matériel développé expérimentalement est présentée.