Considerations on the structure and application of cluster analysis in biology

By **Liviu Dragomirescu**

Extrait des <<Travaux d'Museum d'Histoire naturelle Grigore Antipa >>
VOL. XXVIII
Bucarest, 1986

CONSIDERATIONS ON THE STRUCTURE AND APPLICATION OF CLUSTER ANALYSIS IN BIOLOGY

(Note)

LIVIU DRAGOMIRESCU

INTRODUCTION

I submit some personal reflections (marked with small letters) referring to several questions from the literature (marked with figures) on the structure of the cluster analysis and its application in the biological thinking.

- I begin by specifying the goal for the cluster analysis in biology (item 1 and 2).

- Considering the specified goal, I render evident some analogies between (1) the main groups of algorythms within the cluster analysis; (2) the classification according to the classic logic, and (3) the two "structures" udes by the biologist; the taxonomic ordering and the systemic organization (items 3 and 4).

- On the basis of this discussed analogies, I recommended the cluster analysis algorithms proper to each structure (items 5 and 6).

- At item 7, I indicate (from the literature) a "perfect" algorithm (that leads a unique result). For the study of the systemic organization and I refer to an algorithm (improuved by myself) for the study of the taxonomic ordering.

- At item 8, I defined a concept of homogeneity (for the agglomerative algorithms) and I observer that it is a "conjugated" with the concept of cohesion (for divisive algorithms). Consequently, (1) a step can be taken towards a theory of calssification through the existence of a "perfect" agglomerative algorithm besides a "perfect" divisive algorithm and through a linking bridge between cohesion and homogeneity and (2) two new algorithms can be built.

- At the final item (9), I tried to indicate the biological fields in which the proposed algorithms would be adequate.

THE GOAL AND UTILITIES OF CLUSTER ANALYSIS (IN BIOLOGY)

1. One knows today the relative specialization of the two cerebral hemispheres: the left one is the centre of the discrete algebrais, *algorithmic* thinking and the right one, the centre of the continous, geometric, *heuristic* thinking. The left hemisphere is considered to be under concious control while the "mysterious discoveries" are produced in the right hemisphere.

a). Our civilization is "a civilization of the right hand" controlled by the left hemisphere (cf. Dragomirescu, 1980). Accordingly, there is a dominance of the algorythmic (Dragomirescu, Serban, Banarescu, 1984), mainly expressed by the extension of informatics in all fields.

2. Generally, the cluster analysis is considered to render the clasifications objectives.

a). I consider that the cluster analysis "objectives" the specialist's subjectivity (cf. Dragomirescu, Serban, Banarescu, op. cit) or, more explecitly, the subjectivity of the specialists (who selected the characters for describig the taxonomic

units, codified the data, chose the clustering methods, interpreted the results, etc.) – biologists and mathematicians, in the case of biology.

b). Therefore, I believe that *thecluster analysis must achieve by algorithmic means what the great biologists achieve by euristic means.*[1]

3. Clustering algorithms may be *agglomerative* and *divisive*.

a). One observes that the agglomerative algorithms correspond to a *synthetic* approach while the divisive algorithms correspund to an *analytic* approach.

b). Something new is synthetized and something already existing is analysed.

4. Botnarius (1985) considers, concerning the life evolution, the couple "taxonomic order and systemic organization". According to him, order means "the clustering (…) of some elements (…), based on a certain criterion…" and organization is "the constitution of sets of elements, so that the interactions and their functions are subordinate to the essential functions of the whole".

a). The coupling of items 3b and 4 naturally leads to a conclusion that the taxonomic ordering is synthetized (a construction "on the basis of a criterion") and the systemic organization (something really existing) can be submitted to analysis.

5. It is also Botnariuc (op. cit) who considers that "the main ordering criterion is the degree of similarity…"

a). Consequently, the *taxonomic ordering* will be achieved by using *agglomerative algorithms based on inter-taxa* similarities. Thus, phenogrammes can be achieved.

6. It is known that a real similarity coefficient cannot be obtained through entropy, because such a coefficient would produce the groupping of both the most similar and dissimilar taxonomic units.

a). Considering that the informational entropy (which is a measure of diversity) has the same mathematic formula as the thermodynamis entropy (that measures the dezorganization of a system), I recommend the analysis of the *systemic organization* through *divisive algorithms based on entropy*.

These recommendations are synthetised in Fig. 1.

TOWARDS A THEORY OF CLASSIFICATION. NEW ALGORITHMS

7. Watanabe (1969) says: "If a classification is natural, it is unique", giving Mendeleev's table as an example for the classification of the chemical elements. Therefore, in this case of a given problem, one is interested in the existence of an algorithm that would give a uniquie result (that is the same result, no matter the ordering according to which the taxonomic units are introduced in this algorithm).

[1] An attempt on this matter is, in my opinion, the paper of Dragomirescu, Constantinescu, Banarescu, 1985, drawn up on the basis of the material that was euristically studied by P. Banarescu

Algorithms -> | Based on V	agglomerative	Divisive
Similarity	Synthesis of phenogrammes	
entropy		Analyses of Ecological systems

Among the divisive algorithms there is one (based precisely on entropy) that leads to a aunique result and this is Watanabe's algorithm (1969).

a). As concerns the agglomerative algorithms, I do not know, from the literature, an algorithm with a unique result. Therefore, I took the agglomerative algorithm that is the most consistent from a logical point of view, i.e, the Buser algorithm – given by Buser and Baroni-Urbani (1982 – and I improuved it in order to give a unique result, designating it as the "improuved Buser algorithm" – a contribution of Dragomirescu's doctoral dissertation (1986).

Buser and Baroni-Urbani (op. cit) built the algorithm on the "homogeneity" concept as a generalization (at sets) of the similarities between two elements. I have proposed a homogeneity that models the principles of classification in biology given by Beckner in 1959 (cf. Sneath, Sokal, 1973) and which is a generalization of the most "natural" index of similarity, the Jaccard index (Dragomirescu, Constantinescu, Banarescu, op. cit; Dragomirescu, 1985). Therefore, I nammed it "*Jaccard homogeneity*".

8. Watanabe's algorithm is based on the idea that "any divided system looses cohesion" (Watanabe, op. cit), cohesion being thus, according to the author, a set function (denoted with "c"), which is superadditive, i.e.:

$c (A \cup B) \geq c (A) + c (B)$

Watanabe indicates a single cohesion, i.e. the "W interdependence", based on entropy.

a). I observed (see Dragomirescu, 1986) that a "homogeneity" can be defined in the same general way as a set function (denoted with "o"), which is subadditive:

$o (A \cup B) \leq o(A) + o(B)$

i.e., *homogeneity does not grow by clustering*.

b). It can be shown that the Jacard homogeneity that I proposed is a homogeneity that verifies the condition of item 8a.

c). Starting from a theorem of the measure theory[1] one can build a linking bridge between the divisive and agglomerative algorithms: on the basis given

[1] If $\mu \geq 0$ is an additive measure (i.e.) $\mu (A \cup B) = \mu(A) + \mu(B)$ (V) A, B) that can be written as $\mu = \mu + \mu-$, the statement "$\mu+$ is a superadditive measure" is equivalent to the statement "$\mu-$ is a subadditive measure" ($\mu+$ and $\mu-$ are conjugate as againts $\mu+$) (Watanabe op. cit.).

cohesion, one can uniquely define a "conjugate" homogeneity and vice-versa, (Dragomirescu, 1986).

d). An interesting consequence of the previous item is that one can build an *entropic homogeneity* starting from the cohesion on the entropic basis and a *Jaccard cohesion* starting from Jaccard homogeneity (on the basis of a generalized similarity) (Dragomirescu, 1986). Thus, the vacant places in Fig. 1 can be filled with algorithms (see Fig. 2 – the upper places on the first diagonal).

Seeing it in this way, the theory of classification is likely to become possible through the creation and depending of some links between the theory of measure and topology (knowing that a tree, the result of a cluster analysis, is equivalent to an ultrametric structure). One may possibly think of something analogous to the algebraic topology, that treats the interpretations between the topological structures (of continuity) and the algebraic (discrete) ones.

f). But I have reasons to believe that an extensional approach (Dragomirescu, 1986) will not be sufficient, and an intensional approach will be necessary.

THE UTILITY OF THE NEW ALGORITHMS

9. At item 8d I implicitly proposed two new algorithms: (1) "the improuved Buser algorithm" with the entropic homogeneity, and (2) Watanabe algorithm with Jaccard cohesion.

a). I conclude this note by trying to propose types of problems that could be approach by means of these algorithms. Accordingly, algorithm (1) is good, in my opinion, for simulating some ecological systems (or at least for the short term ecological prognosis of tendency and/or more important than the intergroup ones) and I recommend the algorithm (2) for testing the philogenetic trees (the cladogrammes) (see Fig. 2).

*

Obviously, all the statements relative to the application should be tested. I have already made a few successful tests (some of them already published some others still under investigation) but of course, it isn't but a beginning.

	"Perfect" algorithm	
	agglomerative	divisive
	"improuved Buser algorithm"	Watanabe algorithm
similarity	with Jaccard homogeneity	with JACCARD COHESION
	for synthesis of phenogramme	For TESTING OF CLADOGRAMMES
entropy	With ENTHROPIC HOMOGENEITY	With Watanabe cohesion
	For SIMULATION (EVOLUTION of ECOLOGICAL SYSTEM)	For analysis of ecological system

CONSIDERATII ASUPRA STRUCTURII SI APLICARII ANALIZEI GRUPURILOR IN BIOLOGIE

REZUMAT

Se prezinta pe scurt unele elemente pentru o cenceptie unitara si integrata asupra aplicarii in biologie a analizei gruparilor (clasificarii automate, taxonomiei numerice).

Autorul considera ca analiza gruparilor obiectiveaza subiectivitatea biologului. De aici deduce ca analiza gruparilor trebuie sa realizeze algoritmic ceea ce realizeaza euristic marii biologi.

Pentru aceasta recomanda (1) in *analiza* <<organizarii sistemice >> (din ecologie) algoritmul *diviziv*, care produce rezultat unic, al lui Watanabe si la (2) la *sinteza* <<ordinii taxonomice>> (din sistematica) algoritmul *aglomerativ* dat de Buser si Baroni-Urbani si ameliorat de autor. Se prezinta si o <<legatura>> intre cei doi algoritmi, atat de diferiti, precum si unele consecinte teoretice si aplicative interesante pentru biologie.

REFERENCES

BOTNARIUC (N), 1985 – Cu privire la relatiile dintre ordinea taxonomica si organizarea sistematica a materiei vii. Revolutia biologica: 49 – 58, Bucuresti.

BUSER (M.), BARONI-URBANI (C.) 1982 – A Direct Nondimensional Clustering Method for Binary Data. *Biometrics,* **38,** *2*: 351 – 360.

DRAGOMIRESCU (L.), 1980 – Revolutia informatica (III): Gandirea si informatica. *Studentimea Comunista Brasoneana 37: 9.*

DRAGOMIRSCU (L.), 1985 – Some Extension of Buser and Baroni-Urbani's Clustering method. *Biometrical Journal,* (accepted on June 3th).

DRAGOMIRESCU (L.), 1986 – Contributii privind aplicarea taxonomiei numerice in biologie. Teza de doctorat, Centrul de Statistica matematica, Bucuresti.

DRAGOMIRESCU (L.), SERBAN (M.), BANARESCU (P.), 1985 – A Numerical Taxonomy Method adequate to the Biological Thinking. Applications: Watanabe's Example and the Acanthobrama Genus (Pisces, Cyprinidae). *Trav. Mus. Hist nat. "Grigore Antipa",* **27:** 243-265.

DRAGOMIRESCU (L.), SERBAN (M.), BANARESCU (P.), 1984 – Quelques problemes coccernat l'application de la taxonomie numerique en zoologie systematique. *Trav. Mus. Hist. nat "Grigore Antipa"*, **25:** 361-368.

GIUCULESCU (A.), 1985 – Criza matematicii ansambliste si premizele matematicii intensionale. Matematica in lumea de azi si de maine: 139-153., Bucuresti.

SNEATH (A.), SOKAL (R.), 1973 – Numerical Taxonomy: 1-30, San Francisco ???????.

WATANABE (S.), 1969 – Knowing and Guessing: 299-448. New York

Institutul <<Victor Babes>>
Splaiul Independentei 99-100
76201 Bucuresti, Romania