Biom J. **29** (1987) 3, 345-353 „Some Extensions of Buser and Baroni-Urbani's Clustering Method".

SOME EXTENSIONS OF BUSER AND BARONI-URBANI'S CLUSTERING METHOD

L. DRAGOMIRESCU

"Victor Babeş" Institute, Bucharest (Romania)

Summary

We consider that the algorithm of numerical taxonomy focused upon the concept of "homogeneity" proposed by BUSER and BARONI-URBANI (1982) for binary data, satisfies best from all the other algorithms known by us, the conditions of classification in biology formulated by BECKNER in 1959. Therefore the paper proposes and studies homogeneities for binary and for ordered multi-states data, which model explicitly Beckner's conditions. From all these, the homogeneity denoted by h^* has the remarkable quality of generalizing the famous Jaccard's similarity coefficient. Moreover, using h^* in this algorithm "naturally" improved by us, we can achieve the performance of a correct classification of the remarkable example proposed by WATANABE (1969).

Key words: Modelling; Numerical taxonomy; Biological systematics; Similarity coefficient

1. Introduction

In 1959 Beckner (acc. SNEATH and SOKAL,1973) formulated, as regards the classification in biological systematics, the concept of "polithetic group" or "natural class", definable in terms of a set G of properties $f_1,f_2,...f_n$ so:

"(1) Each one (individual) possesses a large (but unspecified) number of properties in G.
(2) Each f in G is possessed by a large number of these individuals and
(3) no f in G is possessed every individual in the aggregate."

The great majority of classification algorithms tries implicitly to model Beckner's conditions. The algorithm proposed by BUSER and BARONI-URBANI (1982) can be distinguished among all the others, since in our opinion it can model completely properly Beckner's conditions. In DRAGOMIRESCU et al. (1985) we described this algorithm thus: "Let a table which has L lines and N columns, representing a set of L OTU-s (*Operational Taxonomic Unities*, acc. SNEATH and SOKAL, 1973) describes in N characters.

(a) A LIST of all subsets of OTU-s is made up, calculating a homogeneity for each subset.

(b) A subset of the LIST which has the maximum homogeneity from among the other subsets of the LIST is considered a cluster.

(c) If the formed cluster is the whole set then the clustering is over, else the subsets which contain only strict parts of the already formed cluster(s) are eliminated from the LIST and the point (b) is applied to the new LIST."

The most fertile idea of this algorithm is, in our opinion, the concept of homogeneity defined for any set of OTU-s. BUSER and BARONI-URBANI (1982) define two homogeneities for binary data, denoted by h_I and by h_{II}. The homogeneity h_{II} treats equivalently the presences, denoted by 1, and the absences, denoted by 0.

In this paper we are interested only in homogeneities which don't consider the multiple zeroes as a homogeneity argument. This condition is satisfied by the homogeneity h_I which is defined by the above mentioned authors thus:

"Given a data set Λ consisting of L OTU-s and N binary attributes, denote by S_n the sum of the n-th attribute ..."

$$(1) \qquad " \ h_I(\Lambda) = \frac{1}{N \cdot L} \sum_{n=1}^{N} s_n \ 0 \le h_I \le 1 \ "$$

The aim of this paper is to present some mathematical properties of some homogeneities proposed in DRAGOMIRESCU et al. (1985). The proposed homogeneities model explicitly Beckner's conditions.

Section 2 presents homogeneities for binary data, from which h^* has the remarkable property of generalizing in a certain way the famous Jaccard's similarity coefficient.

Section 3 defines and presents properties of an homogeneity (denoted by H^*) for ordered multi-states data, namely tables of ordered continuous or discrete characters. H^* is a generalization of h^*, and has the quality of showing that Beckner's conditions rely on a nonbinary logic.

Section 4 presents an example chosen for various purposes. We mean the famous Watanabe's example (1969) which could classified properly in the logic of the problem only by the algorithm proposed by the same author. Adding a supplementary condition to Buser and Baroni-Urbani's algorithm and applying it whith the homogeneity h^* proposed by us, this remarkable example can be also classified properly.

Secton 5 presents concisely some remarks regarding the proposed homogeneities.

2. Homogeneities for binary data

Let be a set A of L OTU-s described by N binary characters in the form of a binary matrix

$$(a_{ij})_{\substack{i=1,2,\ldots,L \\ j=1,2,\ldots,N}}$$

2.1.1 The homogeneity h

We denote:

$$(2) \qquad m_1(A) = \frac{\sum_{i=1}^{L} ((\sum_{j=1}^{N} a_{ij}) / N)}{L}$$

$$(3) \qquad m_2(A) = \frac{\sum_{j=1}^{N} ((\sum_{i=1}^{L} a_{ij}) / L)}{N}$$

Remark 1: m_1 is the arithmetic average of proportions in which each OTU satisfies the set of characters and m_2 is the arithmetic average of the proportions in which each character is satisfied by a set of OTU-s.

Remark 2: m_1 respectively m_2 can express the degree of satisfying the first, respectively the second Beckner's condition.

Definition 1: The value $h(A)=m_1(A) \cdot m_2(A)$ is called the homogeneity h of the set A of OTU-s.

Properties:

The following propositions are obvious.

Proposition 1. $0 \leq h \leq 1$
Proposition 2. Let $L=1$ then $h(A)=1 <=> (a_{ij}=1\ (\forall)\ j=1,2,...,N)$
Corollary 2.1. The homogeneities h for the sets of an OTU differ (for any matrix having at least a 0)
Proposition 3. $m_1 = m_2$
Proposition 4. $m_2 = h_I$ (from formula (1))

*2.1.2 The homogeneity h**

We denoted by $m_1^*(A)$ respectively $m_2^*(A)$ the values obtained by calculating the formula (2) respectively (3), in which N was replaced by N_1 (N_1 being the number of columns vanish non-identically)

Remark 3: m_1^* is the arithmetic average of the proportions in which each OTU satisfies the set of characters vanish non-identically and m_2^* is the arithmetic average of proportions in which each character vanish non-identically is satisfied by the set of OTU-s.

Remark 4: m_1^* respectively m_2^* can also express the degree of satisfying the first, respectively the second Beckner's condition.

Definition 2: The value $h^*(A) = m_1^*(A) \cdot m_2^*(A)$ is called the homogeneity h^* of the set A of OTU-s.

Properties:
The following propositions are also obvious.

Proposition 5. $0 \leq h^* \leq 1$
Proposition 6. If $L=1$ then $h^*(A)=1$
Proposition 7. $m_1^* = m_2^*$
Corollary 7.1. $h^* = (m_2^*)^2$
Proposition 8. $m_2^* \geq m_2\ (= h_I)$
Corollary 8.1 $h^* \geq h$

The following proposition is a more interesting one. Therefore let us remember the definition of the famous Jaccard's similarity coefficient, which we denoted by $J(A)$, where A is a set of two OTU-s.

It is denoted by

a = the number of the pairs (1,1),
b = the number of the pairs (1,0),
c = the number of the pairs (0,1).

Thus

(4) $$J(A) = \frac{a}{a+b+c}$$

Proposition 9. If L=2 then $m_2^*(A) = (1/2) \cdot J(A)+1/2$

Proof: In order to simplify the denotation we gave up writing the argument. Therefore we have to prove that:

(5) $$m_2^* = (J+1) / 2$$

First we shall prove that

Lemma 9.1

$$m_2^* = \frac{b+c+2a}{2\cdot(a+b+c)}$$

Indeed m_2^* according to the definition and considering that L=2, has the form:

(6) $$m_2^* = \frac{\sum_{j=1}^{N}\sum_{i=1}^{2} a_{ij}}{2 \cdot N}$$

But the numerator is the number of 1-*s* of the two lines matrix.(a*ij*). The same number can be obtained adding the number of columns (1,0), namely *b*, to the number of columns (0,1), namely *c* and to the double number of columns (1,1), namely *a*.

N_1 from the denominator represents the number of the columns vanish non-identically.

It can be easily noticed that N_1 = a+b+c and the lemma proved true. Coming back to the proving of the 9-th proposition, the following difference (denoted by *D*) will be calculated:

(7) $$1 - m_2^* = \frac{b+c}{2\cdot(a+b+c)} = D$$

We notice that m_2* can be written thus:

(8) $$m_2^* = \frac{b+c}{2\cdot(a+b+c)} + \frac{2a}{2\cdot(a+b+c)} = D + J$$

The formula (5) Q.E.D results from the equalization of the values D from the expressions (7) and (8).

3. Homogeneities for Multi-States Data

The logician ENESCU (1980) added a condition, also considered specific to the biological clustering, to Beckner's conditions: "a property is satisfied with more or less intensity".

Accordingly, the OTU-s will be described by multi-states characters, each character being ordered.

3.1 *A homogeneity for ordered multi-states data* (H*)

Let be A a set of L OTU-s described by N ordered (continuous or discrete) characters having the form of a matrix (aij) of real positive numbers.

We denoted the maximum values of each character by:

(9) $$\mu_j = \max_{i=1,2,\ldots,L}(a_{ij})$$

and the sum of these maxima by S:

(10) $$S = \sum_{j=1}^{N} \mu_j$$

By these denotations we define:

(11) $$M_1^* = \frac{\sum_{i=1}^{L}((\sum_{j=1}^{N} a_{ij})/S)}{L}$$

(12) $$M_2^* = \left[\sum_{j(\neq k \text{ for which } \mu_k=0)=1}^{N} ((\sum_{i=1}^{L} a_{ij})/(L\cdot\mu_j)\right]/N_1$$

where N_1 is the number of columns vanish non-identically.

Remark 5:

M_1* is the arithmetic average of the degree in which each OTU-s satisfies the set of characters (vanish non-identically). For an OTU the degree of satisfying the set of characters equalizes the sum of the values of its line divided by the sum of the maxima.

M_2* is the arithmetic average of the degree in which each character (vanish non-identically) is satisfied by the set of OTU-s. For a character the degree of satisfying the set of OTU-s equalizes the sum of the values of its columns divided by the maximum value of the character multiplied by the number of OTU-s.

Remark 6:

M_1* respectively M_2* can also express the degree of satisfying the first, respectively the second Beckner's condition.

Definition 3: The value $H^*(A) = M_1^*(A) \cdot M_2^*(A)$ is called homogeneity H* of the set of OTU-s.

Properties:

Proposition 10. $0 \leq H^* \leq 1$

Proposition 11. If L=1 then H*=1

Proposition 12. H* applied to binary matrices works in the same way like h*.

The proofs of these propositions are immediate.

Property 13. The equalization $M_1^* = M_2^*$ is not valid for any A.

Indeed, let us consider the following example:

Example 1.

Let be the set A consisting of 3 OTU-s defined by 2 ordered characters:

10	1
1	0
0	0
10	1

The maxima are The sum of the maxima is $S = 11$.

Let us calculate:

$$M_1^*(A) = \frac{10+1+1}{3 \cdot 11} = \frac{4}{11}$$

$$M_2^*(A) = \frac{11/(3 \cdot 10) + 1/(3 \cdot 1)}{2} = \frac{7}{20}$$

Therefore $M_1^*(A)$ differs from $M_2^*(A)$

The following property is more interesting.

Property 14. The implication $M_2^*(A) < M_2^*(B) \Rightarrow H^*(A) < H^*(B)$ is not true for any set *A* and any set *B*.

Indeed, let us consider the set from the example 1 to be *A* and the set from the next example to be *B*.

Example 2.
Let be set *B* consisting in 6 OTU-s defined by 2 ordered characters:

100	1
1	0
2	0
0	1
0	1
0	1
100	1

The maxima are: 100 1 The sum of the maxima is $S = 101$.

Let us calculate:

$$M_1^*(B) = \frac{100+1+1+2+1+1+1}{6 \cdot 101} = \frac{107}{606}$$

$$M_2^*(B) = \frac{103/(6 \cdot 100) + 4/(6 \cdot 1)}{2} = \frac{503}{1200}$$

It can be noticed that

$M_2^*(A) = 7/20 < 503/1200 = M_2^*(B)$

but

$H^*(A) = 7 / 55 > 53\ 821 / 727\ 200 = H^*(B)$

Deduction 14.1. By using Buser and Baroni-Urbani's algorithm the homogeneity H^* can produce a classification that differs from that one produced by M_2^* by using the same algorithm.

4. A Remarkable Numerical Example

WATANABE (1969) imagined an example from social psychology, presented in form of a matrix that can appear, in our opinion, in ecoethology too.

Example 3 (WATANABE).

"Suppose that four girl-students live in a dormitory. Three of them are bound by a peculiar mixture of friendship and jealousy, so the group wants to sit alone in the lounge without another member, yet none wants to sit there with both of the remaining two because she cannot stand seeing the evidence of friendship between these latter two. The fourth is entirely neutral to these three and sits in the lounge no matter who else may or may not be sitting there; reciprocally these pay no attention to the fourth girl. Suppose that x_1, x_2, x_3 and x_4 represent these four girls, and yj stands for the predicate 'is sitting in the lounge at the j-th observation':

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
x_1	1	1	0	0	1	1	0	0
x_2	1	1	1	1	0	0	0	0
x_3	0	0	1	1	1	1	0	0
x_4	1	0	1	0	1	0	1	0

This example that, as any pair of OTU-s contains the same number of couples (1,1), (1,0), (0,1), or (0,0), there is no algorithm based upon the similarities able to identify the group of three first students. As far as we know up to now only the algorithm proposed by the same WATANABE (1969) realized the correct grouping of three students.

In the following lines we'll show out that Buser and Baroni-Urbani's algorithm with an improvement proposed in DRAGOMIRESCU et al. (1985) can also realize the performance of grouping the first three students, if using the homogeneity h^*.

The improving of the algorithm consists in adding the following condition to the condition (b) from the description of the algorithm (see Introduction):
"(b') and it has a maximum number of OTU-s."

By calculating the homogeneity h^* for all the subsets of OTU-s the following results are obtained:

-the value 1 for any subset of 1 student (acc. to prop. 6)
-the value $(4/6)^2$ for any subset of 2 students
-the value $(4/6)^2$ for the subset of the first 3 students
-the value $(4/7)^2$ for the other subsets of 3 students and
-the value $(4/7)^2$ for the whole set of students.

Applying the improved algorithm, the first 3 students will cluster, then the whole set will cluster.

It can be easily noticed that, if calculating Buser and Baroni-Urbani's homogeneity h_I, the value $(1/2)^2$ is obtained for any subset. Thus, the improved algorithm will produce the clustering of all the students at a time, and the original algorithm will cluster arbitrarily 2, 3 or 4 students.

5 Discussion

The proposed homogeneities try the explicit modelling of Beckner's conditions and of Enescu's condition.

The homogeneity H^* defined for multi-states data, obviously ordered within each character, shows (acc. to properties 13, 14 and deduction 14.1) that the first two Beckner's conditions are independent, if the problem is expressed nonbinarily. In case of binary data, the first two Beckner's conditions are equivalent, according to the proposition 3 and 7, but we preferred the homogeneities h and h^* instead of their square roots, for comparability with H^*.

In fact the paper treated two homogeneities: Buser and Baroni-Urbani's h_I (or h from item 2.1) and H^* defined at the item 3.1. The first homogeneity is defined only for binary data, and the second homogeneity is defined for ordered multi-states data, therefore also for binary data for which it works identically with h^* from the item 2.2.

We consider as very important, not only principally (a) but also for its applications (b) the fact that the square root from H^* applied to binary data on subsets of 2 OTU-s is a linear function of Jaccard's similarity coefficient (acc. to prop. 12 and 9): (a) Jaccard's coefficient is the most "equilibrated" similarity coefficient which excludes the double zeroes as similarity argument and (b) it is the most frequently in applications. Thus they who use it can have an intuitive view on this homogeneity if knowing how it works when it is particularized as a similarity coefficient.

The reader interested in the value of the homogeneity H^* in application can refer to the paper DRAGOMIRESCU et al (1985) which presents an interesting numerical experiment on the data of the fish genus Acanthobrama, originally codified by Constantinescu, based on Bănărescu's description of the genus.

Acknowledgement

The author is thankful to Prof. Dr. T. POSTELNICU for his guidance.

References

BUSER, W.B., C. BARONI-URBANI, 1982: A Direct Nondimensional Clustering Method for Binary Data. Biometrics 38, 351-360.

DRAGOMIRESCU, L., V. CONSTANTINESCU, P. BANARESCU, 1985: A Numerical Taxonomy Method Adequate for the Biological Thinking. Application: The Acanthobrama Genus and Watanabe's Example. Trav. Mus. Hist. Natt. "Grigore Antipa". Bucarest. 27, 243-265.

ENESCU, Gh., 1980: Fundamentele logice ale gândirii. Editura ştiinţifică şi enciclopedică. Bucureşti.

SNEATH, P.H.A., R.R. SOKAL, 1973: Numerical Taxonomy – the Principles and Practice of Numerical Classification. W.H.Freeman. San Francisco

WATANABE, S., 1969: Knowing and Guessing. John Wiley. New York.