BOL. SOC. BIOL. DE CONCEPCION, TOMO XLIV, pp. 191-200, 1972

UN PROGRAMA DE FORMACION DE AGRUPACIONES APLICADO A LA TAXONOMIA NUMERICA

POR

MIGUEL RAMIREZ * Y J. J. DUEK **

ABSTRACT.

The paper intends to present a basic approach to the problem of obtaining classifications for sets of elements caracterized by families of attributes.

Special emphasis is given to the previous problems of determining the attributes and their states.

A Fortran program is mentioned, implemented in a small size computer, that uses a similarity coefficient and a clustering algorithm previously published.

The technique for computing the similarity coefficient was changed to handle conditioned attributes as tree structures.

INTRODUCCION.

En una gran variedad de aplicaciones es necesario ordenar elementos de acuerdo a las relaciones que presentan los atributos que los caracterizan.

Desde el punto de vista matemático el problema se divide en dos aspectos: 1) la definición y computación de una función de afi-

* Centro de Ciencias de Computación e Información, Universidad de Concepción, Concepción, Chile.
** Departamento de Botánica, Instituto Central de Biología, Universidad de Concepción, Concepción, Chile. Casilla 1367.

nidad o distancia y 2) la construcción y procesamiento de un algoritmo de formación de agrupaciones. De la elección de esta función y algoritmo surge una primera fuente de diferencia en las clasificaciones resultantes, sin embargo quizás más importante que éstas, son las diferencias provocadas por la elección de los elementos del conjunto a clasificar y los atributos que los caracterizarán.

Dentro de la Biología las clasificaciones son parte fundamental de la Sistemática y la Ecología y de hecho buena parte del desarrollo de los métodos numéricos de la clasificación se han efectuado dentro de la Taxonomía Numérica.

Sin embargo, quizás debido a su base matemática, estas materias no han tenido una adecuada difusión.

Con el presente trabajo se busca clarificar los conceptos básicos y presentar una aproximación operacional a la Taxonomía Numérica.

Un problema fundamental ya desde los orígenes de la Taxonomía, ha sido descubrir atributos que permiten caracterizar a los individuos y delimitar los estados en que éstos se presentan, es decir, describir los individuos por la configuración que presentan sus atributos. En una segunda etapa es necesario establecer la estructura en que se ordenan los individuos, de acuerdo a las relaciones de similaridad, que presentan los esquemas con que configuran en los atributos.

La primera etapa es un proceso informado netamente por la experiencia y el conocimiento biológico, en tanto que la segunda, es más bien un problema de simplificación de datos complejos y cae por tanto dentro del tratamiento de información.

1. *Problemas de información biológica*

1.1. *de los individuos*

Puesto que los individuos, más propiamente elementos, a clasificar pueden ser de diverso nivel taxonómico (específico, infraespecífico o supraespecífico) se conviene en designarlos genéricamente por OTU (Operacional Taxonomic Unit).

1.1.1. *Delimitación del conjunto de OTU's a clasificar*

El conjunto de OTU's a clasificar estará condicionado, en primer lugar, por el objetivo del estudio en cuestión. En la práctica, para la mayor parte de los casos, será algún taxa de alguna clasificación disponible la que dé un primer marco para delimitar el conjunto de OTU's a clasificar. Sin embargo es conveniente considerar siempre un conjunto más amplio que aquel en que centremos estrictamente nuestro interés original, ello porque ya que la taxonomía numérica, al basarse en criterios de clasificación, diferentes a los tradicionalmente usados, puede variar la composición y el esquema de orden dados en la clasificación considerada *a priori*. Aún cuando estos

cambios no sean en la mayoría de los casos muy substanciales, el trabajar en un conjunto ampliado, nos posibilita el ubicar nuestro núcleo de interés dentro de un marco de referencia en lo que respecta a sus relaciones de similaridad con taxa próximas.

1.1.2. *Representatividad de los OTU's*

Como ha señalado Mayr, 1968, p. 21, "La substitución del pensamiento tipológico por el pensamiento en términos de poblaciones tal vez sea la máxima revolución conceptual que se haya verificado en biología".

Puesto que, aún a nivel de especies, éstas se componen — en el tiempo y en el espacio — de numerosas poblaciones locales y cuyos individuos difieren entre sí al nivel alélico, parece indispensable el construir los OTU's a partir de una media muestral de estas poblaciones.

Quizás esta sea la parte más larga y costosa del proceso de taxonomía numérica. En ocasiones será incluso impracticable la aplicación de este principio, entonces será necesario apelar a todo el conocimiento que se tenga sobre el OTU a fin de construir un adecuado esquema general de sus atributos.

1.2. *de los atributos*

Consideramos atributos cualesquier afirmación que puede hacerse sobre los OTU's considerados. Según el principio Adansoniano todos los atributos tienen igual valor clasificatorio, es decir conllevan igual cantidad de información y por ende no podemos considerar a unos más importantes que otros. Considerando el atributo como una afirmación que se hace sobre el OTU, esta afirmación puede al menos tener dos sentidos, positivo o negativo, presencia o ausencia de la propiedad observada; en otros casos la afirmación podrá hacerse en más de dos sentidos. En general llamaremos estados de un atributo a los diversos sentidos en que se haga la afirmación respectiva.

1.2.1. *Número de atributos necesarios para caracterizar adecuadamente a los OTU's.*

Este es un problema difícil de precisar.

Puesto que nuestro fin es obtener una clasificación, es decir, en general una relación de orden para el conjunto considerado, el número de atributos necesario será función de la complejidad del sistema (sistema = componentes estructurales y funcionales de los elementos del conjunto + relaciones entre los componentes).

Se supone por lo tanto que mientras mayor número de atributos consideremos, tanto mayor será la precisión de la clasificación obtenida. Un buen índice entonces para encontrar el número buscado, es ob-

servar las variaciones que se producen en la clasificación al agregar nuevos atributos, cuando lleguemos al punto en que no se producen cambios en la clasificación al agregar un nuevo atributo, o éstos son mínimos, consideraremos que el conjunto de OTU's está suficientemente bien caracterizado por los atributos considerados.

Operacionalmente es recomendable considerar de partida el mayor número de atributos que sea factible determinar.

1.2.2. *Atributos que deben evitarse*

Aquellos atributos invariantes, es decir, los que presentan un solo estado para toda la muestra, solo contribuyen a aumentar el nivel general de similaridad; a la inversa los atributos dispersos, esto es, aquellos que presentan un estado diferente para cada OTU del conjunto a clasificar, bajan el nivel general de similaridad. En ambos casos no aportan información que contribuya a agrupar los OTU, y por ende no tiene objeto el considerarlos en la clasificación.

Tampoco son deseables atributos unidos por una relación de dependencia directa pues en realidad constituyen una forma de ponderación de la información que ellos representan.

Los atributos escogidos deben ser aplicables si no a toda la muestra al menos a una gran parte de ella, igualmente no es aconsejable utilizar atributos cuya configuración no sea posible determinar para un número considerable de OTU's.

1.2.3. *Dependencia de atributos*

Hay atributos cuya configuración condiciona la aplicabilidad de otros atributos. Llamaremos a los primeros atributos originales y a los segundos condicionales. Evidentemente existen diversos grados de condicionamiento y así un atributo B, que esté condicionado a un C, y es por lo tanto original con respecto a este último, puede a su vez estar condicionado por un atributo A, y en este caso, podemos considerar a C como en un segundo grado de condicionamiento.

Es necesario considerar esta condición de dependencia entre atributos a fin de evitar la llamada paradoja de Kendrick, que consiste en que OTU's que coinciden en atributos condicionados, pero no en los originales, pueden computarse como más cercanos entre sí que con otros OTU's, con los cuales coinciden en atributos originales, pero no en los condicionados. Esta paradoja se evita si consideramos que los atributos condicionados contribuyen solo a precisar la afinidad de los atributos originales que los condicionen; en consecuencia el atributo original conlleva siempre más información que todos los atributos que él condicione, para ello ponderamos su aporte de información por un factor igual al número de atributos que condiciona directamente (considerando además que cada atributo se condiciona a sí mismo).

Esto no destruye el postulado Adansoniano de igual valor a todos los atributos puesto que la afinidad general entre OTU's se computa solo entre atributos originales, los que se consideran todos en las mismas condiciones, solo que la similaridad de cada uno se ha calculado con las consideraciones ya expuestas.

1.2.4. *Determinación de los estados de un atributo*

Los atributos pueden ser básicamente de dos naturalezas: si hay un número finito de sentidos en que puede expresarse la afirmación, entonces el atributo tiene una distribución discreta y es posible asimilar un estado a cada sentido, o si es necesario reducir el número de estados, agrupando varios sentidos en cada estado. Si en cambio el número de sentidos en que puede expresarse la afirmación no es finito, entonces el atributo tiene una distribución continua y siempre es necesario particionar el intervalo en un número finito de subintervalos, pues de otra manera nos toparíamos con toda seguridad con un atributo disperso, del que ya hemos dicho que no sirve a nuestros fines. Esta distribución continua es una característica de los atributos obtenidos por mediciones.

La partición del intervalo de valores en este último caso, y la agrupación de sentidos en el primero, son problemas que envuelven un considerable grado de subjetividad. Existen tres aproximaciones básicas al problema: 1) buscar como límites de los intervalos las discontinuidades naturales que presenta el esquema de configuración del conjunto de OTU's a clasificar; 2) particionar el intervalo en subintervalos de igual longitud y 3) aplicar alguna distribución estadística en la longitud de los subintervalos.

El objetivo es, en todos los casos, que los elementos de cada subintervalo sean lo más próximos posibles entre sí, y lo más distantes posibles de cualesquier otro en un subintervalo distinto.

1.2.5. *Estados "no aplicable" y "desconocido"*

Los atributos condicionados son pertinentes solo a los OTU's que configuran en ciertos estados del atributo original (precisamente aquellos estados que condicionan su existencia).

Agregamos al atributo condicionado un estado que llamaremos "no aplicable" donde configurarán aquellos OTU's para los cuales el atributo no es pertinente.

El estado "no aplicable" es distinto, y en general tiene un tratamiento diferente, a la condición de desconocimiento de la configuración de un OTU, en un atributo determinado. En este caso consideramos que el OTU configura en un estado que llamaremos "desconocido" cuyo contenido de información para el cómputo de la similitud es siempre nulo.

2. *Problemas de Procesamiento de Información*

2.1. *Elaboración de un coeficiente de similaridad*

El coeficiente de similaridad es una función que asigna a cada par de OTU's un valor real proporcional a la relación que existe entre las configuraciones de sus atributos.

Existen varios coeficientes de similaridad ya establecidos, entre ellos, hemos escogido para trabajar el desarrollado por Rogers y Tanimoto (1960). Este coeficiente se computa como la razón entre el número de estados en que coinciden ambos OTU's y el total de estados en que ambos configuran.

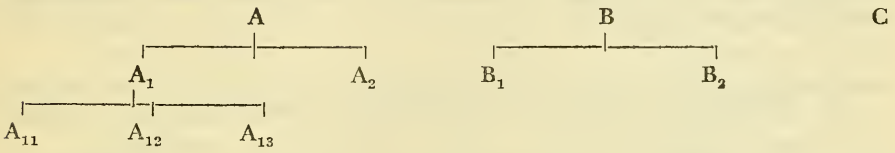
El algoritmo de cálculo del coeficiente se ha modificado para tratar los atributos condicionados.

Los atributos tienen la estructura de un árbol, usando la terminología de estas estructuras diremos que la coincidencia de los atributos-raíz (1.2.3. atributos originales) se computa como un coeficiente de similaridad, para lo cual se consideran los valores de coincidencia de los respectivos subárboles (1.2.3. atributos condicionados) y la coincidencia en el propio atributo-raíz ponderada por un factor igual a su grado (número de subárboles de la raíz) más uno. Podemos por lo tanto construir un algoritmo recursivo de cálculo que vaya del nivel más bajo del árbol hasta su raíz, en lo que podríamos llamar un recorrido post-fijado.

Es decir, se parte calculando la coincidencia entre los atributos de mayor grado de condicionamiento, y los valores obtenidos, se utilizan para calcular la coincidencia entre los atributos del anterior grado de condicionamiento, y así sucesivamente, hasta llegar a obtener los valores de coincidencia de los atributos originales, a los cuales se aplica entonces la fórmula para obtener el valor del coeficiente de similaridad entre los OTU's. La coincidencia entre atributos se calcula según la misma fórmula del coeficiente de similaridad, sólo que ponderamos el aporte de cada atributo, tanto en estados compartidos como en total de estados, por el factor correspondiente a los atributos que él condiciona.

El coeficiente en su forma original solo considera dos posibilidades en la coincidencia de atributos, es decir, si para un atributo los estados en que configuran ambos OTU's coinciden, se suma uno al total de estados compartidos, y uno al total de estados en que ambos configuran; si por el contrario los estados no coinciden, entonces no se adiciona nada al total de estados compartidos y se suma dos al total de estados en que ambos configuran. Con la introducción de los atributos condicionados como modificadores, la coincidencia entre atributos puede tomar cualesquier valor entre 0 y 1, al computar luego el coeficiente de similaridad, debemos sumar el correspondiente valor calculado, al total de estados compartidos, y su complemento con respecto a 2, al total de estados en que ambos OTU's configuran.

Supongamos por ejemplo dos OTU's X e Y, caracterizados por los atributos A, A_1, A_2, A_{11}, A_{12}, A_{13}, B, B_1, B_2, y C, en que las líneas del esquema siguiente, ilustran las relaciones de condicionamiento entre ellos.



Factor	3	4	1	1	1	1	3	1	1	1
Atributo	A	A_1	A_2	A_{11}	A_{12}	A_{13}	B	B_1	B_2	C
Configuración de X	+	+	—	+	—	—	+	+	—	—
Configuración de Y	+	+	+	—	—	+	+	+	+	—
Estados Compartidos	1	1	0	0	1	0	1	1	0	1
Total Estados	1	1	2	2	1	2	1	1	2	1

Coincidencia en A_1 =

$$\frac{\text{coinc. en } A_1 * 4 + \text{coinc. en } A_{11} + \text{coinc. en } A_{12} + \text{coinc. en } A_{13}}{\text{Est. en } A_1 * 4 + \text{est. en } A_{11} + \text{est. en } A_{12} + \text{coinc. en } A_{13}} = \frac{1 * 4 + 0 + 1 + 0}{1 * 4 + 2 + 1 + 2} = \frac{5}{9}$$

Coincidencia en A =

$$\frac{\text{coinc. en A} * 3 + \text{coinc. en } A_1 + \text{coinc. en } A_2}{\text{Est. en A} * 3 + \text{est. en } A_1 + \text{est. en } A_2} = \frac{1 * 3 + \frac{5}{9} + 0}{1 * 3 + \frac{5}{9} + 2} = \frac{16}{29}$$

Coincidencia en B =

$$\frac{\text{coinc. en B} * 3 + \text{coinc. en } B_1 + \text{coinc. en } B_2}{\text{Est. en B} * 3 + \text{est. en } B_1 + \text{est. en } B_2} = \frac{1 * 3 + 1 + 0}{1 * 3 + 1 + 2} = \frac{2}{3}$$

Coeficiente de similaridad X — Y =

$$\frac{\text{coinc. en A} + \text{coinc. en B} + \text{coinc. en C}}{\text{Est. en A} + \text{est. en B} + \text{est. en C}} = \frac{\frac{16}{29} + \frac{2}{3} + 1}{\frac{42}{29} + \frac{4}{3} + 1} = \frac{193}{329}$$

Hay que recordar que el estado "desconocido" no aporta información, por lo tanto no se suma ni a los estados coincidentes ni al total de estados. La situación no es tan clara para el estado "no aplicable" ya que dependiendo de la situación concreta puede ser considerada como un estado normal o bien como otra forma del estado "desconocido".

2.2. *Aplicación de un método de formación de agrupaciones*

La formación de agrupaciones es una forma de simplificar la presentación de las relaciones de similitud, calculadas para el conjunto OTU's, ya que la matriz de datos no nos permite visualizarlas con facilidad.

Estamos limitados a representar estas relaciones en un modelo bi ó, a lo más, tridimensional, y puesto que la función relaciona vectores "n" dimensionales, debemos entrar a alguna forma de compromiso, en la elaboración del modelo.

El método más usual es el llamado de "single link", es decir, un OTU pasa a formar parte de una agrupación al nivel de mayor similaridad, en que se relacione con cualesquiera de los OTU's que ya forma parte de la agrupación. Similarmente dos agrupaciones se fusionan al nivel de mayor similaridad en que dos cualesquiera de sus OTU's se relacionen. Evidentemente entonces el modelo nos muestra el mayor nivel de similaridad con que se relacionan sus elementos.

El otro extremo sería el método del "complete link", es decir, los elementos del modelo se relacionan al nivel de menos similaridad que exista entre dos OTU's de ellos.

El algoritmo que se ha implementado en nuestro programa es el propuesto por Rijsbergen, van (1970), que produce agrupaciones estratificadas jerárquicamente, por el método de "single-link" y una técnica aglomerativa.

3. *Conclusión*

Se ha implementado un programa de Taxonomía Numérica de amplia aplicación, y que contempla la mayor parte de las sofisticaciones conocidas. Este programa ha sido desarrollado en lenguaje FORTRAN para un computador de pequeño tamaño (IBM 1620, 40 K de memoria) por lo que puede ser aprovechado en muchos pequeños Centros de Procesamiento de Datos.

RESUMEN

La taxonomía se ocupa de describir los individuos por medio de sus atributos y luego, utilizando las relaciones de similaridad entre estos atributos, determinar la estructura en que ellos se ordenan.

Para los efectos de representatividad en la descripción de los taxa se estima que los individuos existen en conjuntos de poblaciones locales y por ende debe considerarse una media muestral de ellos.

El número de atributos necesario para caracterizar adecuadamente los OTU's a clasificar es función de la complejidad del sistema considerado. Los atributos invariantes y los dispersos no aportan información útil a la clasificación, luego deben evitarse.

Se presentan tres alternativas básicas para la delimitación de estados en atributos de distribución continua.

Del condicionamiento entre atributos resultan estructuras de árbol. Se considera que los subárboles solo contribuyen a precisar el grado de coincidencia entre los atributos-raíz, por tanto estos últimos conllevan más información que todo el conjunto de sus sub-árboles.

Se propone una técnica de procesamiento basada en un recorrido post-fijado del árbol para determinar la coincidencia entre atributos-raíz, enseguida la similaridad entre OTU's se calcula según el coeficiente de Rogers y Tanimoto y considerando solo la coincidencia de los atributos-raíz.

La estructura del conjunto de OTU's se determina mediante un algoritmo propuesto por van Rijtergen que produce agrupaciones estratificadas jerárquicamente, por el método de single-link.

Todo el proceso ha sido implementado en un programa Fortran para un computador de pequeño tamaño.

SUMMARY

Taxonomy deals with the description of elements by means of their attributes, and then using the similarity relations between these attributes to determine the hierarchical structure they configurate.

To the purpose of representativity in the description of the taxa it is estimated that the element exist in sets of local populations, and so, a sample mean of them must be considered.

The number of attributes necessary to characterize adecuatly the OTU's to classify, is function of the complexity of the taken system. The invariable and the dispersed attributes do not contribute useful information to the classification, so they must be avoided.

There are three basic alternatives for the delimitation of states in continuous distribution attributes.

From the conditioning between attributes arises three structures. If is considered that the sub-trees contribute only to determine precisely the degree of coincidence between the root-attributes, wherefore the latter carry more information than all the set of their subtrees.

A processing technique is proposed, based in a post-fixed run over of the tree, to determine the coincidence between root attributes, then the similarity between OTU's is calculated according to the Rogers & Tanimoto coefficient, considering only the coincidence of the root attributes.

The structure of the set of O.T.U.s determinated by means of an algorithm proposed by van Rijtergen that produce agrupations hierarchicaly stratificated, by the single-link method.

The complete process has been implemented in a FORTRAN program, for a small size computer.

BIBLIOGRAFIA

MAYR, E.
1968 Especies animales y Evolución. España, Univ. de Chile, 808 p.

RIJSBERGEN, C. J. van
1970 Algorithm 52. A fast hierarchic clustering algorithm. The Computer Jour. 13(3) : 234-236.

ROGERS, D. J. et Tanimoto, T. T.
1960 A Computer program for classifing plants. Science 132 (3434) : 1115-1118.

SOKAL, R. R. et SNEATH, P. H. A.
1963 Principles of numerical taxonomy. San Francisco, Freeman, 359 p.

WILLIAMS, W. T.
1969 The problem of attribute-weighting in numerical classification. Taxon 18 (4): 369-374.