Una nota sobre el procedimiento en Nuevos algoritmos no-jerárquicos en clasificación de datos de López-Hernández

por

ARACELI GARÍN
Dpto. Economia Aplicada III
Fac. CC. EE. y Empresariales
Universidad del Pais Vasco
Avd. Lehendakari Aguirre 83, 48015 Bilbao
Tfno:94-6013734
Email:etpgamaa@bs.ehu.es

Resumen

En esta breve nota presentamos una crítica al procedimiento empleado en el denominado Algoritmo Inductivo en los Objetos de López-Hernández(1997). El objetivo de dicho algoritmo es la clasificación de un conjunto de objetos en subconjuntos o clusters maximales, o equivalentemente, la determinación de los ciclos maximales en el grafo de proximidad constituído por el conjunto de objetos para un nivel de proximidad dado. Aunque su elevada complejidad lo clasifica como un problema *NP*-duro, desde los años sesenta y setenta han sido desarrollados diversos algoritmos computacionalmente eficientes. En el caso del procedimiento propuesto, hay una ausencia total de justificación tanto metodológica como computacional.

Palabras Clave: cluster maximal, ciclo, subgrafo completo, grafo no dirigido.

Clasifciación AMS: 62H30, 68Q25.

1. INTRODUCCION

L.J. López y M. Hernández (1997) presentan dos nuevos algoritmos no jerárquicos para clasificación de datos. El objetivo del primero de ellos es recubrir el conjunto de los *n* objetos $O = \{o_i\}_{i=1}^{n}$ con una familia de clusters maximales.

Definición 1

Una medida de desemejanza *D,* es una aplicación $D: OxO \to R^{+} \cup \{0\}$, que satisface:

1) $D(o_i, o_i) = 0, \forall i = 1,...,n.$ $_i$

2) $D(o_i, o_j) = D(o_j, o_i), \forall i, j \in \{1,...,n\}, i \neq j.$

Se dice que D es una distancia si además satisface la desigualdad triangular:

$$D(o_i, o_j) \leq D(o_i, o_k) + D(o_k, o_j) \forall i, j, k \in \{1,...,n\}, i \neq j \neq k.$$

Definición 2

Sea $O = \{o_i\}_{i=1}^{n}$ el conjunto de objetos y *D* una función de desemejanza sobre *O.* Dado un nivel de desemejanza ε >0, una clase o cluster sobre *O* es un subconjunto *C* de *O,* tal que $\forall o_i, o_j \in C, D(o_i, o_j) \leq \varepsilon.$

Definición 3

Dado el conjunto de objetos $O = \{o_i\}_{i=1}^{n}$ y el nivel de desemejanza ε >0, se define el grafo de proximidad $G_\varepsilon(V, A)$, donde el conjunto de nudos o vértices es *V=O,* o alternativamente $V = \{i \in \{1,...,n\} : o_i \in O\}$ y el conjunto de arcos $A = \{(i, j) \in V \times V : D(o_i, o_j) \leq \varepsilon\}$.

Entenderemos entonces, que dos vértices cualesquiera $(i, j) \in V \times V$ están relacionados si existe un arco que los une. La relación anterior, extendida a un subconjunto $(o_1,...,o_k)$ de *O* nos dice que dichos elementos están ralacionados si y sólo si existe un arco entre cada uno de los elementos o_i y todos los demás $o_j, j = 1,...,k, j \neq i.$ En este caso $(o_1,...,o_k)$ constituye un subgrafo completo de cardinalidad *k* o equivalentemente un cluster.

Definición 4

Se dice que un subgrafo completo es maximal si no está contenido en ningún otro subgrafo completo. El cluster correspondiente a un subgrafo completo maximal, se denomina cluster maximal.

Un grafo no dirigido $G = (V, A)$ con $|V| = n$, puede ser representado por una matriz simétrica de orden $n \times n$, $M = (m_{ij})$, denominada matriz de adyacencia de G, donde $m_{ij} = 1$ indica la existencia de arco entre los nudos o_i y o_j; $m_{ij} = 0$, en otro caso.

2. MOTIVACION

A la vista de las anteriores definiciones, el problema de clasificación de objetos del conjunto O en clusters maximales para un nivel de desemejanza dado ε >0, se resuelve al obtener el conjunto de subgrafos completos maximales en el grafo no dirigido $G_\varepsilon(V, A)$.

La determinación de subgrafos completos en un grafo no resulta en absoluto un problema novedoso. Es un problema que se platea en muy diversas aplicaciones, siendo en las décadas de los años 60 y 70 cuando más algoritmos se publican en relación a este tema. (Johnson (1957), Moon y Moser (1965), Auguston y Minker (1970), Peay (1970), Bron y Kerbosch (1973), Leifman (1976), o Gerhards y Lindenberg (1979), entre otros).

La alta complejidad del problema a resolver, se trata de un problema *NP*-duro, ha llevado a desarrollar procedimientos que incluyen fundamentalmente metodologías branch-and-bound.

Todos los algoritmos para determinar el conjunto de subgrafos completos en un grafo finito, $G = (V, A)$, operan a partir de la matriz de adyacencia. Estos algoritmos pueden ser divididos en dos clases: la primera de ellas incluye a los algoritmos que proporcionan un mayor ahorro de memoria, ya que no precisan guardar ninguna información de cada subgrafo completo maximal después de su obtención. La otra clase contiene aquellos algoritmos que construyen sucesivamente cada subgrafo completo maximal de G, pero necesitan tener almacenadas largas listas de subgrafos completos no maximales durante su ejecución. En la primera de las clases citadas se encuentra el algoritmo de Bron-Kerbosch (1973) (Algoritmo 457 de la CACM), considerado hoy en día como el más eficiente. Publicada su implementación en Algol 60, dicho procedimiento puede ser fácilmente traducido a C (o FORTRAN). De esta forma consiguen algunos de ellos una eficiciencia, en media, cercana a la óptima posible.

Por otra parte, es habitual que el desarrollo de un nuevo algoritmo lleve consigo una justificación de la metodología empleada, así como resultados computacionales, realizados sobre conjuntos de grafos de distinto tamaño y densidad (normalmente generados aleatoriamente), con el fin de mostrar la eficiencia del mismo, o comparar su tiempo de ejecución con el de alguno de los ya existentes.

3. CONCLUSIONES

En esta nota se critica el procedimiento descrito como Algoritmo Inductivo en los Objetos López-Hernández (1997). No tratándose de un problema novedoso y existiendo herramientas con eficiencia contrastada, capaces de resolver el problema planteado, no parece evidente la necesidad de desarrollo de una nueva.

Si aún conociendo la existencia de algoritmos factibles para el problema, el objetivo ha sido desarrollar uno nuevo, hubiera sido interesante ver descrita la metodología empleada, así como ver contrastada su eficiencia con alguno de los ya existentes, o incluso, la variación de su tiempo de ejecución a medida que aumenta la dimensión del problema.

La posibilidad de implementación de cualquiera de las herramientas existentes, nos ha permitido replicar la tabla correspondiente y detectar una errata en el Ejemplo 1 de aplicación descrito en las

páginas 134 y 135. Concretamente para $\varepsilon = 32$, el primer cluster maximal debería ser $\{o_2, o_4, o_5, o_6, o_9\}$ en lugar del allí obtenido: $\{o_2, o_4, o_5, o_6\}$.

En ningún caso, la crítica realizada trata de cuestionar la metodología propuesta en López-Hernández (1997), como técnica de clasificación no jerárquica adecuada en determinados contextos.

4. AGRADECIMIENTOS

La autora agradece la financiación al proyecto de investigación 038.321-HA129/99 de la UPV.

Referencias

Auguston, J. G. y Minker, J. (1970). *An analysis of some graph theoretical cluster techniques.* Journal of ACM, 571-588.

Bron, C. y Kerbosch, J. (1973). *Finding all cliques of an undirected graph*. Communications of the ACM, 16, n. 9, 575-577.

Gerhard, L. y Lindenberg, W. (1978). *Clique detection for nondirected graphs: two new algorithms*. Computing, 21, 295-322.

Johnson, L.F. (1957). *Determining cliques on a graph*. Proc. of the fifth Conference on Numerical Mathematics, 429-437.

Leifman, L.J. (1976). *On construction of all maximal complete subgraphs (cliques) on a graph*. Preprint. Dept. of Mathematics, University of Haifa, Israel.

López, L.J. y Hernández, M. (1997). *Nuevos algoritmos no jerárquicos en clasificación de datos*. Estadística Española, 39, n. 142, 129-140.

Moon, J.W. y Moser, L. (1965). *On cliques in graphs*. Israel J. of Mathematics, 3, 23-28.

Peay, E.R. (1970). *An iterative clique detection procedure*. Michigan Math. Psychology Program: MMPP, 70-74.