EVALUACIÓN DE HEURÍSTICAS DE OPTIMIZACIÓN COMBINATORIA EN CLASIFICACIÓN POR PARTICIONES*

Alexia Pacheco Hernández[1], Programa de Posgrado en Matemática, Universidad de Costa Rica, Costa Rica
Javier Trejos[2], Eduardo Piza[3] y Alex Murillo[4],
CIMPA, Universidad de Costa Rica, Costa Rica, 2060 San José, Costa Rica

RESUMEN
El propósito de este artículo es presentar los resultados de la evaluación de heurísticas de optimización combinatoria por particiones, específicamente, sobrecalentamiento simulado, búsqueda tabú y algoritmos genéticos, en comparación con métodos tradicionales como k-medias y clasificación jerárquica de Ward. Se utilizaron tablas de datos generadas al azar de acuerdo con ciertos parámetros establecidos. Se generaron 16 tablas de datos con variables normalmente distribuidas, se repitió el experimento 100 veces para cada tabla y cada método, y como parámetro de comparación de los resultados se utilizó la inercia intra-clases (W). Los mejores resultados se obtuvieron para el sobrecalentamiento simulado y el algoritmo genético.

ABSTRACT
The aim of this paper is to present the results of the evaluation of combinatorial optimization heuristic applied to obtain partitions in clustering: simulated annealing, tabu search and a genetic algorithm, using data tables generated randomly according to some defined parameters. Those techniques were compared between them and with traditional methods (k-means and Ward's agglomerative clustering). Sixteen tables were generated with normally distributed variables and for each one, the experiment was repeated 100 times for each method. The intra-classes inertia was used as criterion to compare the classifications obtained. Best results were obtained for simulated annealing and the genetic algorithm.

Key words: simulated annealing, taboo search, genetic algorithms, k-means, hierarchical clustering, simulation.

MSC: 90C59 562H30 90-08

1. INTRODUCCIÓN

Varios métodos tradicionales en Análisis Multivariado de Datos encuentran óptimos locales de los criterios que optimizan. Tal es el caso en clasificación por particiones, escalamiento multidimensional, regresión no lineal y conjuntos burdos.

Recientemente, muchas heurísticas de optimización han sido propuestas; su finalidad es encontrar óptimos globales en problemas de optimización discreta. Entre estas heurísticas, están el sobrecalentamiento simulado (SS) [8], la búsqueda tabú (BT) [3] y los algoritmos genéticos (AG) [4].

Muchos autores han tratado de encontrar mejores soluciones a los problemas de Análisis Multivariado de Datos utilizando estas heurísticas de optimización. En la Universidad de Costa Rica el equipo PIMAD del Centro de Investigación en Matemática Pura y Aplicada (CIMPA) ha abordado distintos problemas: clasificación numérica, clasificación binaria, clasificación bimodal, escalamiento multidimensional métrico, escalamiento unidimensional, regresión no lineal, selección de variables, y conjuntos burdos. Se han obtenido resultados aplicando SS, BT y AG sensiblemente mejores a los obtenidos con los métodos tradicionales, en tablas de datos ya conocidas y probadas en diferentes literaturas, como son: i) la tabla de Datos escolares [15]

*Este trabajo fue presentado en el 6to. Taller Internacional de Investigación Operacional celebrado en el periodo 7-11 Marzo, 2005 en Ciudad Habana.

E-mail: [1]alexiapachecohernandez@hotmail.com
[2]epiza@cariari.ucr.ac.cr
[3]jtrejos@cariari.ucr.ac.cr,
[4]murillof@cariari.ucr.ac.cr

(9 objetos y 5 variables), ii) los Peces de Amiard [1] (23 de objetos, 16 variables), iii) la Sociomatriz de Thomas [1] (24 objetos, 24 variables), iv) los Iris de Fisher [2] (150 objetos, 4 variables).

En el presente trabajo se hace una evaluación comparativa entre cinco métodos de clasificación automática por particiones, en el caso de datos numéricos, utilizando técnicas de simulación tipo Monte Carlo.

En la sección 2 se recuerda el problema de la clasificación numérica, con el criterio de inercia o varianza intraclases a minimizar. La sección 3 explica brevemente los métodos que se emplearán en la comparación. La sección 4 indica los objetivos del experimento y la sección 5 explica cómo se procedió a la generación de las tablas de datos y en ella se presenta un resumen de los resultados obtenidos. Finalmente, la sección 6 contiene las conclusiones del trabajo.

2. CLASIFICACIÓN NUMÉRICA

El objetivo de la clasificación automática (conocida como *clustering* en inglés) es encontrar grupos homogéneos de objetos, de tal forma que objetos similares pertenezcan a la misma clase, y que sea posible distinguir entre objetos que pertenecen a clases diferentes.

En el caso numérico, se tiene el conjunto de objetos $\Omega = \{x_1, x_2, \ldots, x_n\}$ tal que $x_i \in R^p$, para todo índice i, esto es, los objetos son descritos por p variables numéricas o cuantitativas. El criterio más ampliamente usado es la minimización de la varianza o inercia intraclases:

$$W(P) = \frac{1}{n} \sum_{l=1}^{k} \sum_{x_i \in C_l} \| x_i - g_l \|^2, \qquad (1)$$

donde K es el número (fijado de antemano) de clases, $P = (C_1, C_2, \ldots, C_k)$ es la partición que se busca, y g_k es el centro de gravedad o vector promedio de C_k. Minimizar W(P) es equivalente a maximizar la varianza o inercia interclases:

$$B(P) = \sum_{l=1}^{k} \frac{|C_l|}{n} \| g_l - g \|^2, \qquad (2)$$

donde g es el centro de gravedad total y $|C_k|$ es la cardinalidad de la clase C_k , ya que la suma W(P) + B(P) es una constante (la inercia total).

Debe observarse que este criterio satisface la propiedad de monotonicidad:

$$\text{Min} \{ W(P'): P' \in P^*_{k+1} \} \leq \min \{ W(P) : P \in P^*_k \}, \qquad (3)$$

donde P^*_k es el conjunto de todas las particiones de Ω en exactamente k clases no vacías. Esto significa que no tiene sentido comparar particiones con diferente número de clases, y por ello el número de clases es fijado de antemano.

3. LOS MÉTODOS A EVALUAR

K-medias es el método de clasificación automática más conocido en la literatura, el cual iterativamente va mejorando una partición dada, mediante el cálculo de los centros de gravedad de las clases y la asignación de los objetos a la clase cuyo centro es más cercano; se repite el procedimiento hasta obtener estabilidad en la partición. Este es un caso típico de búsqueda local y es bien sabido que la solución obtenida depende de la partición dada inicialmente.

La clasificación jerárquica aglomerativa construye árboles de clasificación o dendrogramas, de acuerdo con un criterio de agregación. El criterio de Ward se aplica generalmente a datos numéricos y minimiza el incremento de la inercia. Tiene la propiedad de no producir inversiones. Para particionamiento, se corta el árbol jerárquico en el número de clases deseado para la comparación.

La aplicación de heurísticas modernas de optimización combinatoria -como es el caso de SS, BT y AG- está basada en el uso de transferencias de objetos de una clase a otra. En el caso del sobrecalentamiento

simulado (SS) [5, 6, 8, 12] se usa la regla de Metrópolis para decidir si transferencias de un objeto a una clase (ambos escogidos al azar) se efectúa. En búsqueda tabú [3, 9, 10] se construye una serie de particiones (el vecindario de una partición dada) por la transferencia de un único objeto a la vez, y se escoge la mejor partición según las reglas de esta técnica. Finalmente, se aplica un algoritmo genético [4, 12, 16] con una representación cromosómica de n alelos en un alfabeto de K letras que representa a una partición, y se usan los operadores de selección proporcional a B(P), mutaciones (que corresponden a transferencias), y un cruzamiento especial que hemos llamado "cruzamiento forzado". En todos los tres casos anteriores, el uso de heurísticas mostró un comportamiento claramente superior al de las k-medias o cualquier otro método conocido de particionamiento al aplicarlo a tablas de datos ya conocidas en la literatura.

Con el fin de evaluar el comportamiento de tales heurísticas y de compararlas con algoritmos tradicionales, se diseñó un experimento que se relaciona a continuación.

4. OBJETIVO DEL EXPERIMENTO

El objetivo del experimento es comparar los siguientes métodos de clasificación automática:

- Sobrecalentamiento simulado (SS)
- Búsqueda Tabú (BT)
- Algoritmo genético (AG)
- Nubes dinámicas o k-medias (KM).
- Clasificación jerárquica (agregación de Ward) (WARD)

Se estableció como parámetro de comparación la mínima inercia intraclases W(P) y el porcentaje de veces que se alcanza ese valor, después de realizar 100 réplicas de la aplicación del método a cada una de las tablas de datos.

5. TABLAS DE DATOS SIMULADOS

Se generaron tablas con generadores de números semi-aleatorios siguiendo la distribución normal multivariada, considerando 4 factores y dos niveles en cada uno de ellos. Los factores son:

- El número n de individuos; se tomó n = 105 y n = 525.
- El número K de clases; se tomó K = 3 y K= 7.
- La cardinalidad de las clases; en el primer nivel se tomaron todas las clases con una misma cardinalidad, mientras que en el segundo nivel se tomó una clase mayor que el resto (con aproximadamente el 50% de todos los objetos, y en las clases restantes se distribuyen equitativamente los demás objetos).
- La varianza de las clases; en el primer nivel se tomaron todas las clases con varianza igual a uno, y en el segundo nivel una clase con el triple de la varianza que el resto de las clases.

En todos lo casos las tablas tienen p = 6 variables, normalmente distribuidas. Los vectores de medias fueron generados al azar en $[0,1]^6$. Por lo tanto, se tiene 16 casos, y para cada uno se generaron 100 particiones iniciales al azar antes de aplicar los métodos de particionamiento. En este experimento se puede medir el porcentaje de mala clasificación, por la forma en que se construyeron las tablas de datos.

En SS se usaron los siguientes parámetros: tasa inicial de aceptación de particiones que empeoran el criterio χ_0 = 95 %, tasa de decrecimiento geométrico de la temperatura γ = 0,9, longitud de las cadenas de Markov $L_t = 100n(K - 1)$, número de máximo de iteraciones t_{max} = 150, límite para detener las cadenas de Markov = 10n(K - 1) y m = 4 como valor máximo permitido de iteraciones que repiten el mismo valor final.

La BT empleó los siguientes valores |T| =7 como longitud de la lista tabú y un máximo de iteraciones de $t_{max} = nK|T|$.

El AG se usó con los siguientes parámetros: tamaño de población M = 40, probabilidad de cruzamiento p_c = 0.25 y de mutación p_m = 0.001, número máximo de iteraciones $t_{max} = MK\log(n)$ y frecuencia para aplicar nubes dinámicas = 10. La partición inicial generada al azar que utilizan todos los otros métodos que lo requieran se toma como el primer miembro de la población; los restantes 39 miembros se generan al azar.

La Tabla 1 muestra un resumen de los resultados de las 100 corridas para cada método.

Con base en los resultados presentados en la Tabla 1 se ve que el método de Ward tiende a obtener menor calidad que los métodos de particionamiento cuando las desviaciones estándar son diferentes; de lo contrario, sus soluciones son comparables.Cuando las cardinalidades de las clases son iguales, los métodos de particionamiento obtienen los mismos óptimos; sin embargo, el SS es claramente superior, con una tasa de atracción del mejor valor de W del 100%.

Cuando la cardinalidad de las clases es diferente y se buscan muchas clases, el SS y la BT tienen grandes dificultades, con resultados más pobres que para k-medias. En este caso, el AG es claramente superior a los otros métodos, aún si no siempre encuentra el óptimo.

Tanto por la tasa de atracción como por el valor del criterio, el AG es mejor que los restantes métodos en este experimento.

Debe hacerse notar que, a pesar de no estar reportados los tiempos de computación, el método de k-medias es mucho más rápido que los demás, y que el más lento es el BT. En promedio, en el peor de los casos, este último tardó unos 16 minutos por corrida, el AG 2 minutos, el SS unos 30 segundos, y el k-medias muy pocos segundos.

Tabla 1. Resultados de aplicar sobrecalentamiento simulado (SS), búsqueda tabú (BT), algoritmo genético (AG), k-medias (kM) y clasificación jerárquica de Ward en las 16 tablas de datos; se reporta el mejor valor de W y el porcentaje de veces que ese factor fue encontrado en 100 corridas.

n	k	SS		BT		AG		kM		Ward
		W	%	W	%	W	%	W	%	W
		Cardinalidades iguales								
		Varianzas iguales								
105	3	5.422	100	5.422	99	5.422	100	5.422	91	5.42
105	7	5.146	100	5.146	74	5.146	82	5.146	19	5.15
525	3	5.993	100	5.993	100	5.993	100	5.993	98	5.99
525	7	5.339	100	5.339	82	5.339	88	5.339	45	5.34
		Varianzas diferentes								
105	3	13.15	100	13.15	99	13.15	100	13.15	13	13.85
105	7	09.8951	100	9.895	51	9.895	69	9.895	1	10.17
525	3	15.809	100	15.809	51	15.809	82	15.81	2	16.41
525	7	8.261	100	8.261	100	8.261	94	8.261	53	9.37
		Cardinalidades distintas								
		Varianzas iguales								
105	3	5.007	100	5.007	100	5.007	100	5.007	91	5.01
105	7	6.991	62	6.991	1	5.545	35	5.545	3	5.55
525	3	5.672	8	5.672	100	5.672	100	5.672	96	5.67
525	7	8.105	100	8.105	1	5.648	22	5.648	2	5.66
		Varianzas diferentes								
105	3	11.734	100	11.734	100	11.734	100	11.734	95	11.86
105	7	8.654	100	8.654	40	7.625	37	7.625	6	7.69
525	3	13.819	3	13.819	100	13.819	100	13.819	59	14.2
525	7	8.497	100	8.518	1	7.456	21	7.463	2	8

6. CONCLUSIONES

Según se ha podido apreciar con este experimento, las heurísticas de optimización combinatoria permiten mejorar los resultados de los métodos tradicionales de clasificación automática, en el caso de datos numéricos, sobre todo para el algoritmo genético y el sobrecalentamiento simulado.

Debe notarse que las heurísticas necesitan de un buen generador de números aleatorios, por lo cual el usuario debe ser cuidadoso con la escogencia del generador. Es común que los generadores implementados en los compiladores tengan problemas [13] (es decir, que no pasen los tests de aleatoriedad), lo mismo que muchos paquetes con librerías de rutinas. Nosotros hemos empleado el método sustractivo de D. Knuth [7] y empleando la tabla de números aleatorios de la Rand Corportation [14].

Cabe destacar que para el experimento se generó una plataforma de software en Delphi 6, lo que facilitó la aplicación del mismo. Esta plataforma puede ser empleada para comparar mejoras que se le hagan a las heurísticas aplicadas o incluir nuevas heurísticas.

REFERENCIAS

CAILLIEZ, F. and J.P. PAGES (1976) : **Introduction à l'Analyse des Données**. SMASH, París.

EVERITT, B.S. (1993): **Cluster Analysis**. 3a edición. Edward Arnold, Londres.

GLOVER, F. **et al**. (1993) "Tabu search: an introduction", **Annals of Operations Research** 4(1-4): 1-28.

GOLDBERG, D.E. (1989): "Genetic Algorithms in Search", **Optimization and Machine Learning**. Addison-Wesley, Reading-Mass.

KIRKPATRICK, S.; D. GELATT and M.P. VECCHI (1983): "Optimization by simulated annealing", **Science** 220: 671-680.

KLEIN, R.W. and R.C. DUBES (1990) "Experiments in projection and clustering by simulated annealing", **Pattern Recognition** 22: 213-220.

KNUTH, D.E. (1981): "Seminumerical Algorithms". Segunda edición, volumen 2 del libro **The Art of Computer Programming.** Addison-Wesley, Reading, Mass.

LAARHOVEN, P. and E. AARTS (1988): **Simulated Annealing: Theory and Applications**. Kluwer Academic Publishers, Dordrecht.

MURILLO, A. and J. TREJOS (1996) : "Classification tabou basée en transferts", S. Joly & G. Le Calvé (Eds.), **IV Journées de la Société Francophone de Classification**, Vannes: 26. 1-26.

MURILLO, A. (2000): "Aplicación de la búsqueda tabú en la clasificación por particiones", **Investigación Operacional** 21: 183-194.

PIZA, E. (1987): "Clasificación Automática Jerárquica Aglomerativa", **Revista de Ciencias Económicas** 7(1).

PIZA, E.; A. MURILLO and J. TREJOS (1999): "Nuevas técnicas de particionamiento en clasificación automática", **Revista de Matemática**: Teoría y Aplicaciones 6: 51-66.

PRESS, W.H.; B.P. FLANNERY; S.A. TEULOLSKY and W.T. VETTERLING (1990): **Numerical Recipes**. **The Art of Scientific Computing**. Cambridge University Press, New York.

THE RAND CORPORATION (1955): **A Million Random Digits with 100,000 Normal Deviates**. The Free Press, Glencoe.

SCHEKTMAN, Y. (1978): "Estadística Descriptiva", I Parte, Memorias **I Simposio Métodos Matemáticos Aplicados a las Ciencias,** J. Badia, Y Schektman y J. Poltronieri (eds.), Universidad de Costa Rica, San Pedro: 9-67.

TREJOS, J. (1996): "Un algorithme génétique de partitionnement", S. Joly & G. Le Calvé (Eds.), **IV Journées de la Société Francophone de Classification**, Vannes: 37.1-37.