Revista de Matemática: Teoría y Aplicaciones 2003 **10**(1–2) : 11–22
cimpa – ucr – ccss issn: 1409-2433

MÉTODO HEURÍSTICO PARA PARTICIONAMIENTO ÓPTIMO

Sergio G. de-los-Cobos-Silva* Javier Trejos Zelaya†
Blanca Rosa Pérez Salvador‡ Miguel Angel Gutiérrez Andrade§

Recibido: 14 Enero 2002

Resumen

Muchos problemas en el análisis de datos requieren del particionamiento no supervisado de un conjunto de datos dentro de clases o conglomerados no vacíos que sean bien separados entre ellos y lo más homogéneos entre sí. Un particionamiento ideal es cuando se puede asignar cada elemento del conjunto a una clase sin que exista ambigűedades. Este trabajo consta de dos partes principales; primero se presentan diferentes métodos y heurísticas para encontrar la cantidad de clases en que se debe particionar un conjunto de manera óptima; posteriormente se propone una novedosa heurística y se realizan algunas comparaciones para observar sus ventajas considerando conjuntos muy conocidos y utilizados que están previamente clasificados presentándose al final algunos resultados y conclusiones.

Palabras clave: Particionamiento óptimo, clasificación, heurísticas.

Abstract

Many data analysis problems deal with non supervised partitioning of a data set, in non empty clusters well separated between them and homogeneous within the clusters. An ideal partitioning is obtained when any object can be assigned a class without

*Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana – Iztapalapa, Av. Michoacán y La Purísima s/n, Col. Vicentina, Del. Iztapalapa, México D.F., C.P. 09340 México. Fax: 58.04.46.40. E-Mail: cobos@xanum.uam.mx

†CIMPA, Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica. E-Mail: jtrejos@cariari.ucr.ac.cr

‡Departamento de Matemática, Universidad Autónoma Metropolitana – Iztapalapa, Av. Michoacán y La Purísima s/n, Del. Iztapalapa, México D.F., México. E-Mail: psbr@xanum.uam.mx

§Departamento de Sistemas, Universidad Autónoma Metropolitana - Azcapotzalco, Av. San Pablo No. 180, Col. Reynosa Tamaulipas, Del. Azcapotzalco, México, D.F., C.P. 02200; Fax: (52)53.94.45.34; E-Mail: gama@correo.azc.uam.mx

ambiguity. The present paper has two main parts; first, we present different methods and heuristics that find the number of clusters for optimal partitioning of a set; afterwards, we propose a new heuristic and we perform different comparisons in order to evaluate the advantages on well known data sets; we end the paper with some concluding remarks.

Keywords: Optimal partitioning, clustering, classification, heuristics.

Mathematics Subject Classification: 62H30, 91C20, 90C59, 68W20.

1 Introducción

Muchos problemas en el análisis de datos requieren del particionamiento de un conjunto de datos dentro de clases o conglomerados no vacíos que sean bien separados entre ellos y lo más homogéneos entre sí. Un particionamiento ideal es cuando se puede asignar cada elemento del conjunto a una clase sin que exista ambigüedades. Muchas técnicas se han propuesto para dicho análisis considerando ya sea un esquema de separación bajo la suposición de que los conglomerados tienen una forma hiperesferoidal como es el caso entre otros del uso de la suma de cuadrados. Si la separación de las fronteras entre los diferentes conglomerados no son lineales, entonces los métodos basados en la suma de cuadrados, como es el caso del método de k-medias fallan; por lo que últimamente se han propuesto diferentes esquemas de solucón para tratar de superar este problema, como es en [2] donde se propone el uso de núcleos, transformaciones no lineales del conjuto de datos originales a un espacio característico de mayor dimensión, donde los datos son linealmemte separables y entonces se trabaja directamente en este espacio característico. También se han propuesto métodos que tratan de superar la dificultad anterior como en [5] y [7] mediante el proponer la minimización de la entropía y considerando una mezcla gaussiana de las funciones de densidad de probabilidad. Existen algunas propuestas para clasificación utilizando algún esquema de tipo difuso como en [3].

El esquema propuesto en este trabajo y presentado parcialmente en [1] no tan sólo proporciona el número optimal de conglomerados sino que además nos permite identificar los datos de cada conglomerado y que a diferencia de otros métodos la heurística propuesta en este trabajo (en adelante la denotaremos como **SCA**) no depende de ninguna estimación inicial puesto que, algunos (si no es que todos) los métodos dependen del valor inicial de uno o varios parámetros para estimar el número de conglomerados como en [2] o se requerirá de la selección a priori de la mezcla de gaussianas como en el caso de [5]. SCA se puede considerar como un método de búsqueda dirigida para encontrar el número de particiones "naturales", tratando de escapar de la optimalidad local. En la segunda sección se hace una breve descripción de algunas de las técnicas del Estado del Arte utilizadas en la clasificación. En la tercera sección se presenta la heurística propuesta junto con algunos resultados obtenidos. Por último, se proporcionan algunas conclusiones y perspectivas.

2 Algunos criterios utilizados en el problema de clasificación

2.1 Criterio de la suma de cuadrados

El criterio de la suma de cuadrados es tal vez el más utilizado no tan sólo para el problema de particionamiento sino también como criterio para muchos otros problemas en optimización.

Para el problema de clasificación, la matriz de costos de la suma de cuadrados para un conjunto finito de observaciones, digamos $X = \{x_1, x_2, \ldots, x_n\}$ donde $x_i \in \mathbb{R}^P, i = 1, 2, \ldots, n$ y dados K conglomerados centrados, está dada por:

$$SC = \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n} z_{j,i}(x_i - m_j)(x_i - m_j)^T, \tag{1}$$

donde m_j es el valor medio del conglomerado j y $z_{j,i}$ es una función indicadora que es igual a la unidad si el elemento i está en el conglomerado j y cero de otra forma. La matriz indicadora $Z = [z_{j,i}]_{K\times n}$ donde $z_{j,i} \in \{0, 1\}, \forall j, i$ y $\sum_{j=1}^{K} z_{j,i} = 1, \forall j$, en el caso de que se tenga conglomerados hiperesféricos, proporciona un medio para encontrar el particionamiento optimal del conjunto de datos X, mediante la solución del siguiente problema de optimización:

$$Z^* = \min_{Z} \text{Traza}(SC). \tag{2}$$

Varios métodos como el de las k-medias están basados en esta medida, la cual implica que se tienen conglomerados de forma hiperesférica, por lo que si los datos se clasificaran en conglomerados con una forma geométrica diferente, este criterio no sería el adecuado.

2.2 Espacio característico y uso de núcleos

En [2] se aborda el problema en el que los conglomerados no son hiperesféricos, por lo que se considera entonces un mapeo no-linear, suave y continuo del espacio de datos X hacia un espacio característico digamos F tal que:

$$\phi : \mathbb{R}^D \to F, \tag{3}$$

por lo que suponiendo K conglomerados centrados y utilizando la notación anterior, la matriz de costos de la suma de cuadrados está dada por:

$$SC^{\phi} = \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n} z_{j,i}(\phi(x_i - m_j^{\phi})\phi(x_i - m_j^{\phi})^T), \tag{4}$$

donde m_j^{ϕ} es el valor medio del conglomerado j para los valores transformados, de manera análoga al caso de mínimos cuadrados, ahora lo que se desea es:

$$Z^* = \min_{Z} \text{Traza}(SC^{\phi}). \tag{5}$$

Recuérdese que un núcleo(kernel en inglés) de valor real definido en R^p es una función digamos $KN : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow R$ la cual es simétrica y definida positiva. En particular son de interés los núcleos que se pueden expresar como producto interno, i.e., $KN(x, y) =< \phi(x), \phi(y) >$.

Bajo la consideraciónes de que se tiene una función de base radial se construye el núcleo $KN_{ij} = KN(x_i, x_j) = exp(\frac{-1}{c}||x_i - x_j||^2)$ y utilizando el teorema de convolución para gaussianas se tiene que:

$$\int_x p(x)^2 dx \approx \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} KN_{ij}, \tag{6}$$

por lo que el problema a resolver es:

$$Z^* = \min_Z \text{Traza}(SC^\phi) = max_Z \sum_{j=1}^{K} \gamma_{jn} R(x|C_{kn}), \tag{7}$$

donde $R(x|C_k) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ki} z_{kj} KN_{ij}$, que proporciona una medida de la compacidad de C_k y $\gamma_{kN} = n_k/n$.

Posteriormente en [2], se considera que los eigenvectores de la matriz KN proporcionan un medio para estimar el número de conglomerados inherentes al conjunto de valores y se propone un procedimiento iterativo para calcular los valores de $Z = [z_{ij}]_{n \times n}$. Cabe mencionar que dependiendo del valor del parámetro de la función de base radial que está utilizando el autor el estimado inicial de conglomerados varía respecto a éste, por lo que el método propuesto no es eficiente.

2.3 Entropía mínima

En [5] se propone una partición mediante la mezcla de gaussianas. Inicialmente se considera que dada una partición del conjunto de datos X en K conglomerados se tiene que la probabilidad de que un dato $x \in X$, condicionada a la partición, está dada por:

$$p(x) = \sum_{k=1}^{K} p(x|k) p(k) \tag{8}$$

y considerando la medida Kullback-Lieber(KL) de sobreposición de dos distribuciones y que está dada por:

$$KL(p(x)|g(x)) = \int p(x) ln(\frac{p(x)}{g(x)}) dx. \tag{9}$$

Entonces lo que se desea es minimizar la sobreposición entre las dos distribuciones. Utilizando el teorema de Bayes se llega a:

$$V = \int H(x) p(x) dx, \tag{10}$$

donde $H(x) = -\sum_k p(k|x) \ln p(k|x)$, es la entropía de Shannon.

En este trabajo el autor no indica el cómo se debe escojer tanto la cantidad de gaussianas como sus respectivos parámetros de manera explícita, por lo que también queda como un ejercicio de aproximación para dar respuesta al problema.

Los mismos autores en un segundo trabajo [7] proponen una mezcla semiparamétrica de funciones gaussianas y consideran que las clases condicionales a posteriori se pueden tomar como una combinación de núcleos a posteori i.e. $p = W\Phi$ donde p es el conjunto de las probabilidades a posteori de las clases en forma vectorial, W es una matriz de transformación y Φ es el conjunto de núcleos a posteriori. Posteriormente consideran un movimiento de $W \longrightarrow W^{'}$ asociado a un cambio en la dimensión intrínseca de $K \longrightarrow K^{'}$ (i.e. un cambio en la hipótesis del número de clases de la partición) y utilizando una modificación de la ecuación de Metropolis-Hastings dada por:

$$p(\text{aceptar}) = \min\{1, \frac{p(K^{'}, W^{'}|\Phi)}{p(K, W|\Phi)}\text{Jacobiano}(W, K \longrightarrow K^{'}, W^{'})\frac{q(K, W|\Phi)}{q(K^{'}, W^{'}|\Phi)}\}, \qquad (11)$$

posteriormente proponen algunos criterios de movimientos, actualización, nacimiento y muerte para la elección del número de conglomerados, lamentablemente en este trabajo no se reportan resultados de IRIS.

3 SCA: heurística propuesta

En esta sección se dan los lineamientos del método heurístico de clasificación propuesto y que en adelante lo referiremos como SCA (sistema de clasificación aleatorio), la cual se expone en la Figura 1. Donde la matriz AGRUPA es una matriz de $n \times n$ que lleva la frecuencia en que los diferentes elementos se agrupan entre sí cada vez que se mejora con respecto a la función objetivo, la cual fue en todos los casos en este trabajo de la forma $exp(\frac{-1}{c}||x_i - x_j||^2)$. Cabe mencionar que se utilizaron varios valores para la constante c en los diferentes ejemplos de este trabajo y en todos los casos se encontraron resultados similares.

En términos generales la idea de SCA se puede parafrasear de la siguiente manera:

1. Dado un conjunto de elementos que tienen alguna(s) característica(s) en común proyéctese ese conjunto a un espacio característico donde sean linealmente separables en particular si se utilizan funciones núcleo(que se puedan expresar como producto punto de funciones) este espacio característico es de igual dimensión que el original.

2. Genere aleatoriamente particiones del conjunto y tome aquellas que mejoran con respecto a la función objetivo, llevando la frecuencia en que cada elemento se agrupa con los demás.

3. Obtenga mediante el procedimiento de la Figura 1 el número de conglomerados y los elementos característicos de cada uno de ellos.

4. Con esta información inicial proceda a refinar el particionamiento utilizando alguno de los métodos descritos en algunas de las referencias, por ejemplo en [8].

1. Proporcione el valor de Iteraciones, una clasificación inicial C_0 de los n datos y calcule el valor de la función objetivo $F_0(C_0)$.
2. Para $i = 1$ hasta Iteraciones:
3. Asigne de manera aleatoria una partición C_i y calcule $F_i(C_i)$
   - Si $F_{i-1}(C_{i-1}) < F_i(C_i)$ entonces actualiza la matriz AGRUPA
   - Fin_Si
4. Fin_Para
5. A partir de la matriz AGRUPA encuentra la matriz Δ de disimilitudes, ver Trejos(1996).
6. A partir de la matriz de disimilitudes encuentre el número de conglomerados y los elementos característicos de cada uno.
7. Para encontar el número de conglomerados, a partir de la matriz de disimilitudes $\Delta = [\delta_{i,j}]$ construya la matriz $X = [X_{i,j}]$, como se explica a continuación:
   - $Xi, j = \frac{1}{n-i+1} \sum_{k=i+1}^{n} \delta k, j.$
   - $Xj, min = \min_j\{X_{i,j}\}$, esto representa el "centraje" de los datos.
   - $Xj, max = \max_j\{X_{i,j}\}$, siendo esta cantidad la "excentricidad" de los datos.
   - Encuentre la media, la varianza y el coeficiente de variación para Xj, min, y para Xj, max.
   - Escoja el coeficiente de variación con la menor varianza como el porcentaje de cada conglomerado y redondee al entero más cercano, este último es el número estimado de conglomerados del conjunto de datos.
8. Identifique los elementos característicos de cada conglomerado a partir de la matriz Δ.

Figura 1: Pasos principales de SCA.

Datos	K	$\hat{K}$	n	dim.
IRIS(3)	3	3	150	4
IRIS(2)	2	2	100	4
PIMA	2	3	768	7
Círculos	2	2	63	2
Conjuntos		6	65	2
Petu	2	2	194	4

Tabla 1: Tabla comparativa donde K = No. de clases originales, $\hat{K}$ = No. de clases estimadas por SCA, n=número de elementos del conjunto y dim = No. de características de cada elto.

3.1 Resultados

En esta sección revisamos algunos conjuntos de datos.

3.1.1 Conjunto IRIS

Un caso clásico muy conocido y utilizado es el conjunto de datos de IRIS. Este ejemplo consta de elementos no linealmente separables(ver la Figura 2). Primero se realizó un estudio sobre todos los datos obteniendo que se tenían tres clases diferentes(ver la Figura 3), aunque dos de las cuales no se diferenciaban muy bien entre sí pero si con respecto de la tercera, como se puede observar en la Figura 3. Posteriormente, se realizó el estudio utilizando SCA con tan sólo los 100 elementos de las dos clases no distinguibles y se encontró que que se trataba de dos clases diferentes(ver la Figura 4).

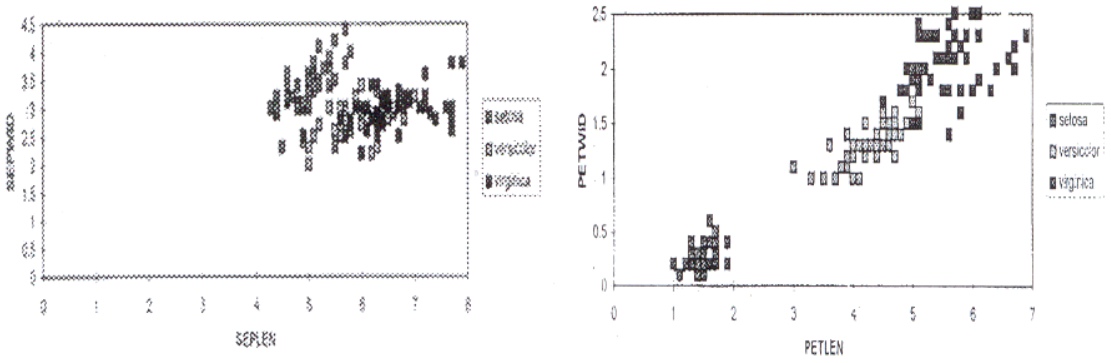

Figura 2: Algunas gráficas de los datos de IRIS.

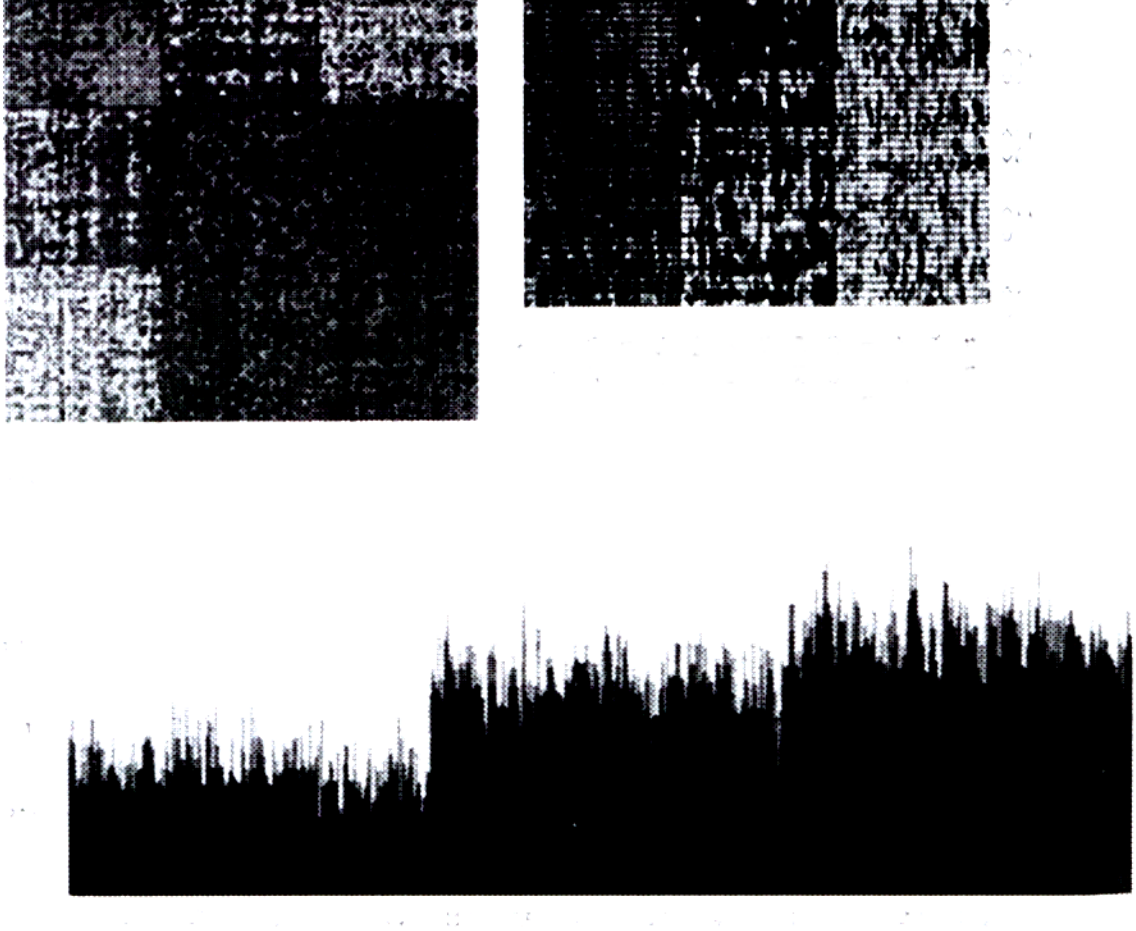

Figura 3: Algunas gráficas de los 150 datos de IRIS al correr SCA.

3.2 Otros ejemplos

Se consideraron otros ejemplos a los que denotaremos como CIRCULOS, GRUPOS, PIMA y PETU. En el ejemplo de PIMA sólo se tomó una muestra aleatoria de 265 elementos, los datos se pueden obtener de [11]. Los datos de PETU se obtuvieron de [4]. Es interesante el observar que es difícil ver en el ejemplo llamado CONJUNTOS(ver parte izquierda de la Figura 6), si los datos en la parte inferior izquierda se trata de uno o dos conjuntos.

4 Conclusiones

En [2], se necesita conocer de antemano el número de conglomerados para posteriormente utilizar el método propuesto por lo que el problema de encontrar el número de conglomerados no está resuelto.

En [5] y [6] para el mismo problema de IRIS propone una mezcla de **20** gaussianas aunque no indica el cómo encuentra los diferentes parámetros de éstas y los mismos autores en [7] proponen un método pero no atacan el problema de IRIS.

Queremos por último indicar que por la experiencia computacional con los ejemplos que se han trabajado tan sólo entre el 4% y 6% del total de las iteraciones proporcionan mejora, en todos los ejemplos presentados se tomaron entre 1000 y 2000 iteraciones de mejora, por lo que una línea de investigación será el cómo mejorar este porcentaje, así como el de utilizar otras funciones objetivo. La tabla de la Tabla 1 nos proporciona una idea de la eficiencia de SCA.

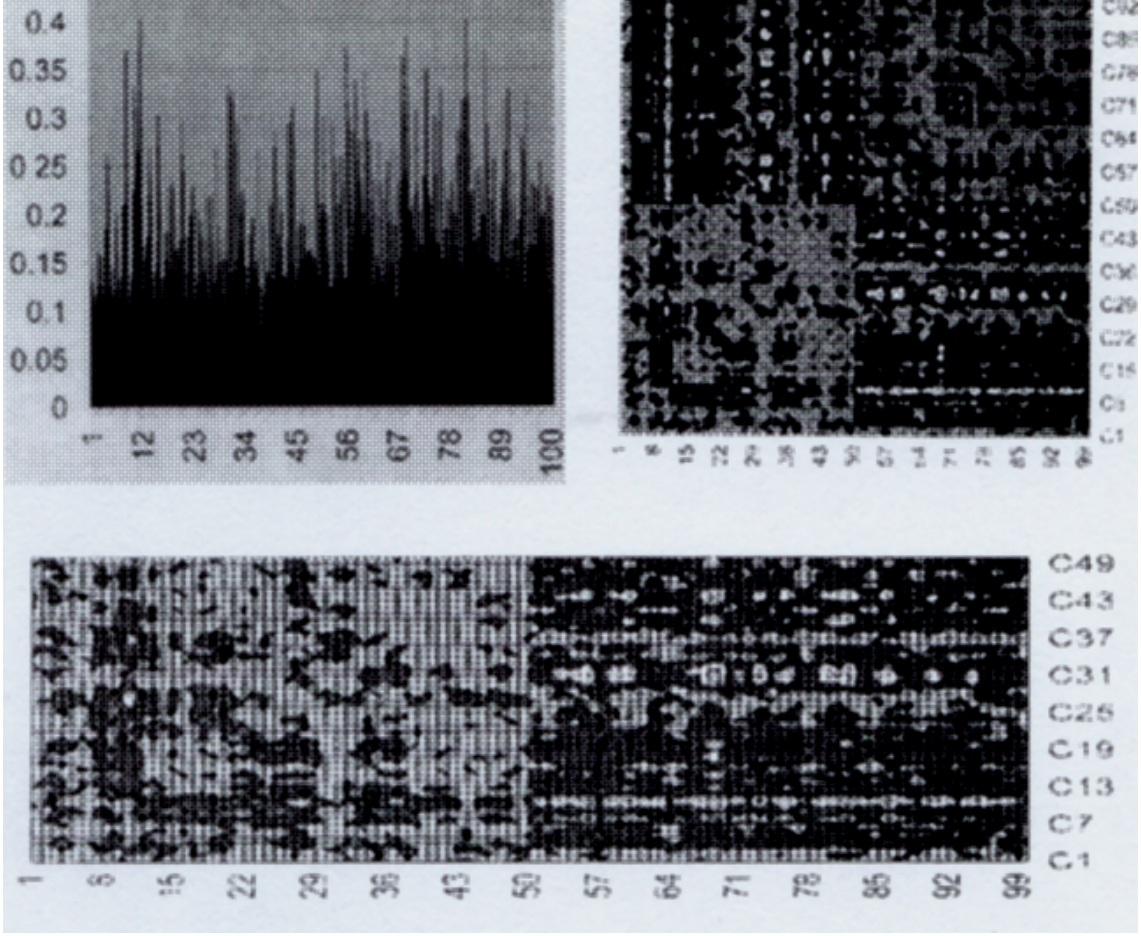

Figura 4: Algunas gráficas de los 100 datos de IRIS no linealmente separables al utilizar SCA.

Referencias

[1] de-los-Cobos-Silva, S.; Goddard, J.; Pérez, B.R.; Gutiérrez, M.A. (2001) "SCA: sistema de clasificación aleatoria", XV Foro Nacional de Estadística, 8-12 de octubre, Guadalajara-México.

[2] Girolami, M. (2001) "Mercer kernel based clustering in feature space", *I.E.E.E. Transactions on Neural Networks* (to appear).

[3] Goddard, J.; de-los-Cobos-Silva, S. (2000) "On a class of distance metrics for fuzzy c-means", *Proc. VII Congress of SIGEF*, Chania, Greece: 577–584.

[4] Goddard, J.; Martínez, A.E.; Martínez F.M. (1998) "Prototype selection for nearest neighbour classification", *Congreso Latinoamericano de Ingeniería Biomédica*, Mazatlán, México.

[5] Roberts, S.J.; Everson, R.; Rezek, I. (2000) "Maximum certainty data partitioning", *I.E.E.E. Patterns Recognotition* **33**(5): 833–839.

[6] Roberts, S.J.; Everson, R.; Rezek, I.(2001) "Minimum entropy data partitioning", Technical report, IISGroup, Dep. EEE, Imperial College of Science Technology & Medicine, U.K.

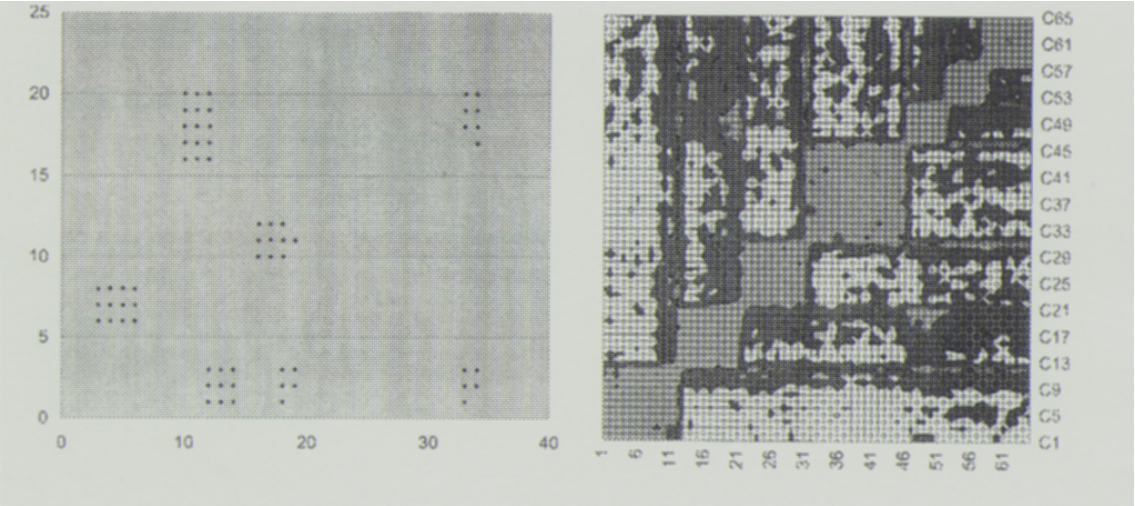

Figura 5: Algunas gráficas de los datos CONJUNTOS al utilizar SCA.

[7] Roberts, S.J.; Everson, R.; Rezek, I.(2001). "Minimum-entropy data clustering using reversible jump Markov chain Monte Carlo", (to appear in *IEEE*).

[8] Trejos, J.; Murillo, A.; Piza, E. (1998) "Global stochastic optimization for partitioning", in: A. Rizzi et al. (Eds.), *Advances in Data Science and Classification*. Springer, Heidelberg: 185–190.

[9] Romesburg, H.C. (1984) *Cluster Analysis for Researchers*. Krieger Publishing Company.

[10] Trejos, J. (1996) "Propiedades y aplicaciones de una medida de redundancia de la información: el número equivalente", Mem. *X Foro Nacional de Estadística* y *II Congreso Iberoamericano de Estadística*, Oaxaca-México: 221–226.

[11] `http://www.ics.uci.edu/pub/machine-learning-databases`.

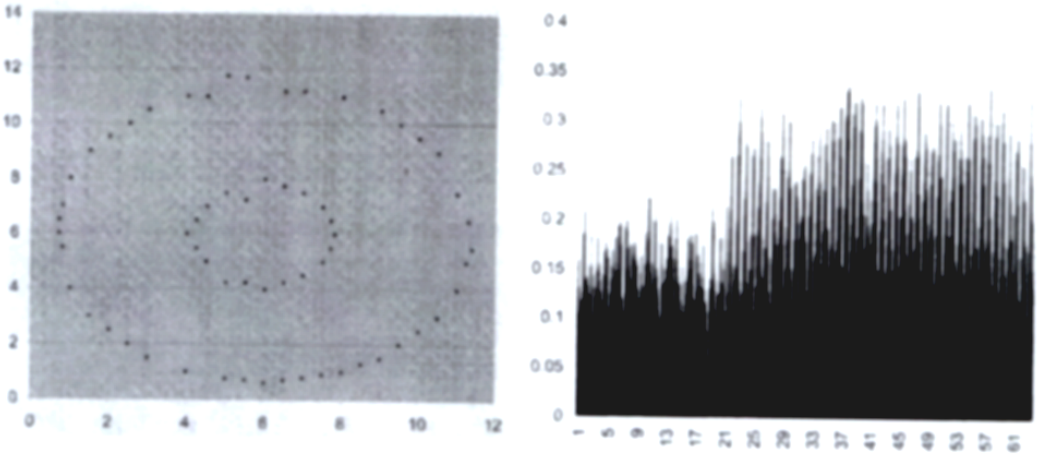

Figura 6: Algunas gráficas de los datos CIRCULOS al utilizar SCA.

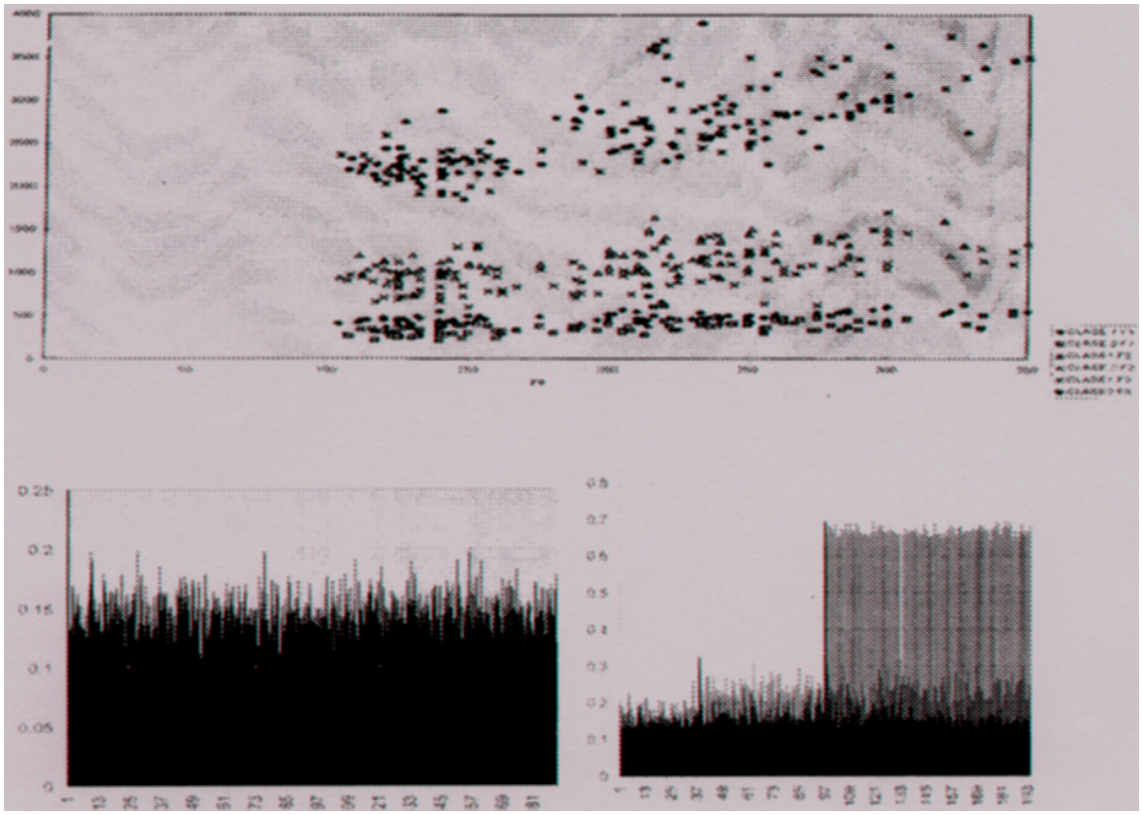

Figura 7: Algunas gráficas de los datos de PETU al utilizar SCA.

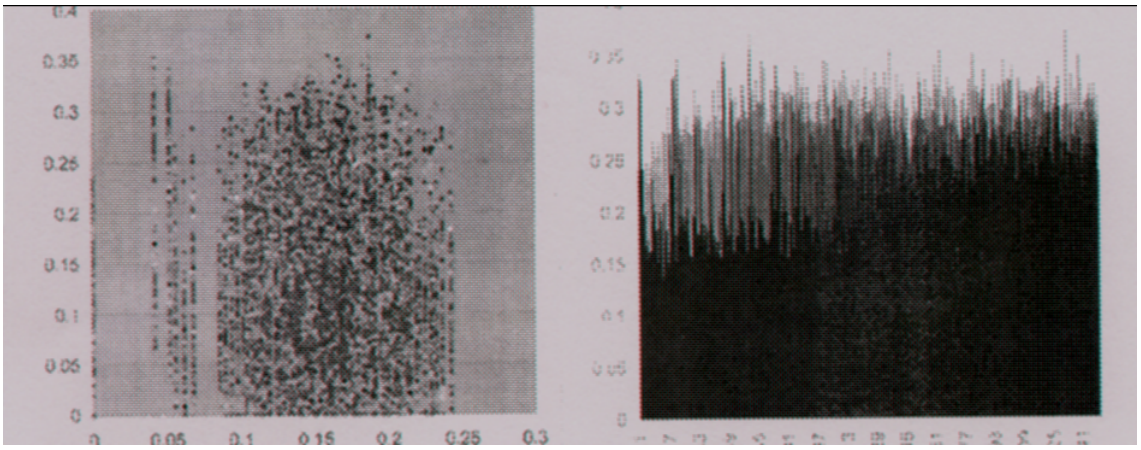

Figura 8: Algunas gráficas de los datos de PIMA al utilizar SCA.