Clustering basado en densidad

Christopher Expósito Izquierdo
Airam Expósito Márquez
Israel López Plata
Belén Melián Batista
J. Marcos Moreno Vega

{**cexposit, aexposim, ilopezpl, mbmelian, jmmoreno**}**@ull.edu.es**

Departamento de Ingeniería Informática y de Sistemas
Universidad de La Laguna

Índice

CLUSTERING BASADO EN DENSIDAD

Clustering basado en densidad

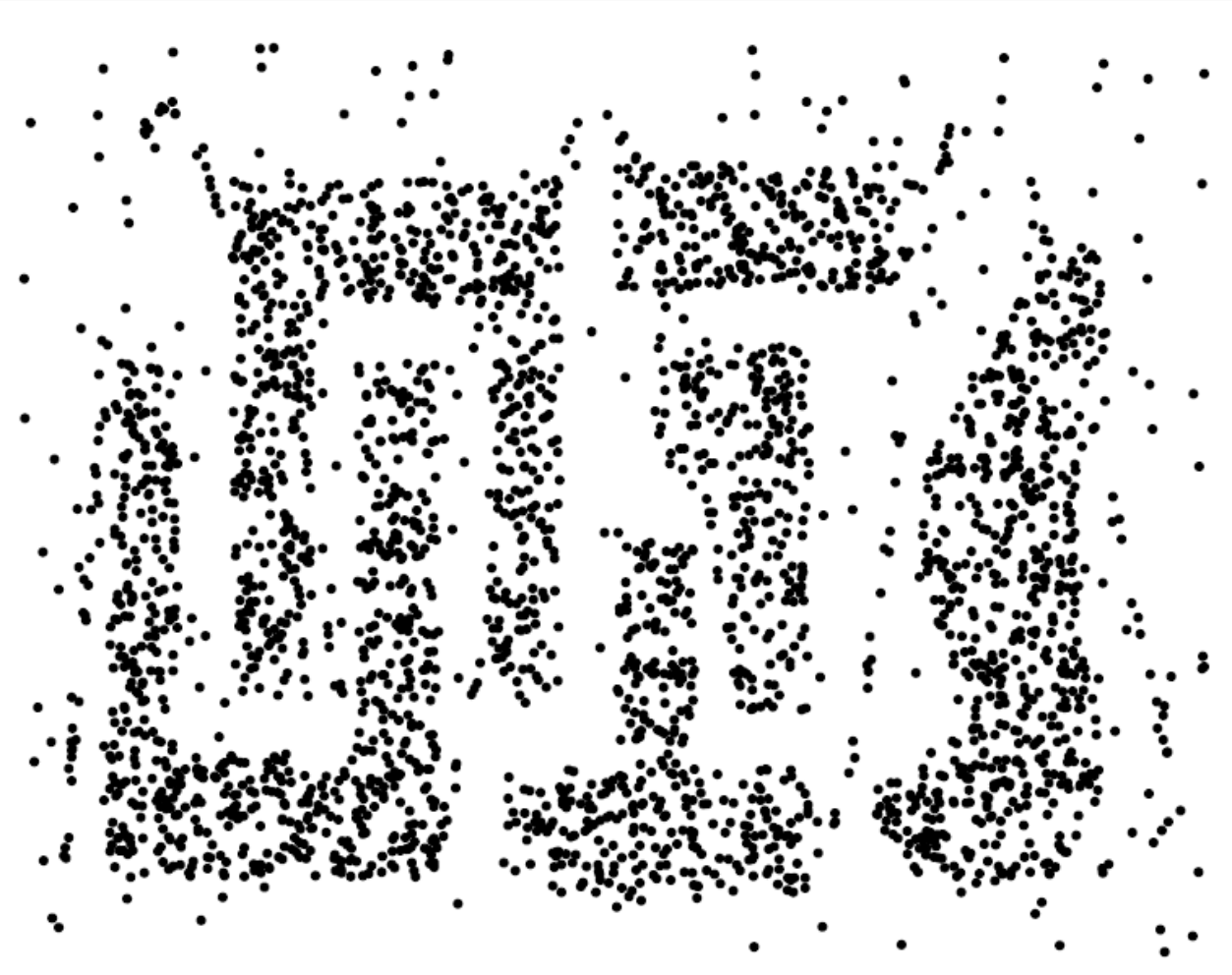

Figura 1: Agrupamiento basado en densidad

Clustering basado en densidad

Idea

- Los algoritmos para clustering basado en densidad identifican regiones de alta densidad que están rodeadas de áreas poco densas.
- Cada una de las regiones densas identificadas se corresponde con un cluster.
- El clustering basado en densidad es apropiado cuando los clusters no tienen una forma geométrica definida.

Algoritmo DBSCAN (i)

- DBSCAN [1] es un algoritmo de densidad simple que implementa la noción de densidad por medio de un procedimiento basado en centro.
- La densidad de cada punto se estima contando el número de puntos, incluido él, que se encuentran a una distancia no mayor que *Eps* del punto.
- Dependiendo de dicho valor, cada puntos es clasificado como central (core), frontera (border) o ruido (noise).

Densidad

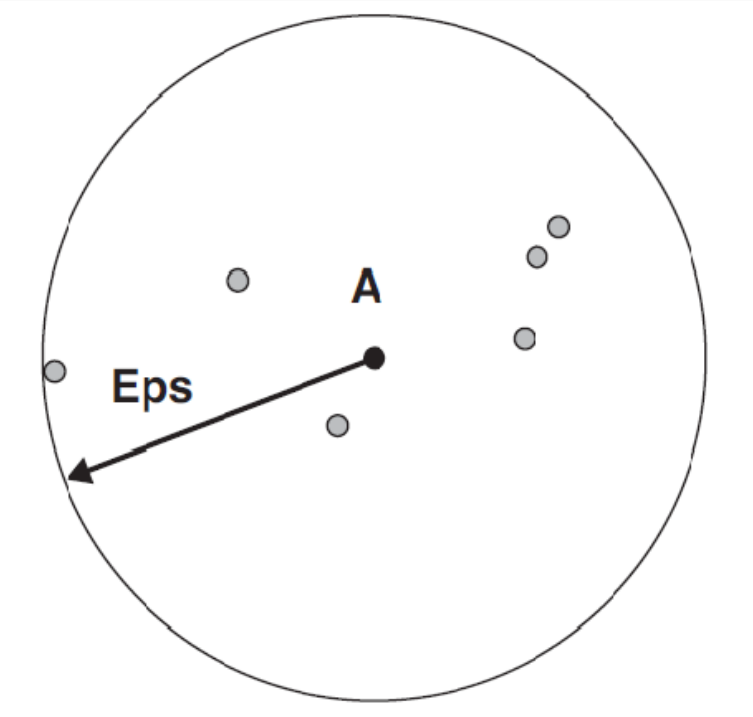

Figura 2: Densidad basada en centro

Clustering basado en densidad

Distintos puntos

- **Punto central.** Un punto es central si el número de puntos a una distancia no mayor que *Eps* de él supera el valor mínimo *MinPts* (parámetro del algoritmo).
- **Punto frontera.** Un punto es frontera si no es central pero hay al menos un punto central a una distancia no mayor que *Eps*.
- **Punto ruido.** Un punto es ruido si no es central ni frontera.

Clustering basado en densidad

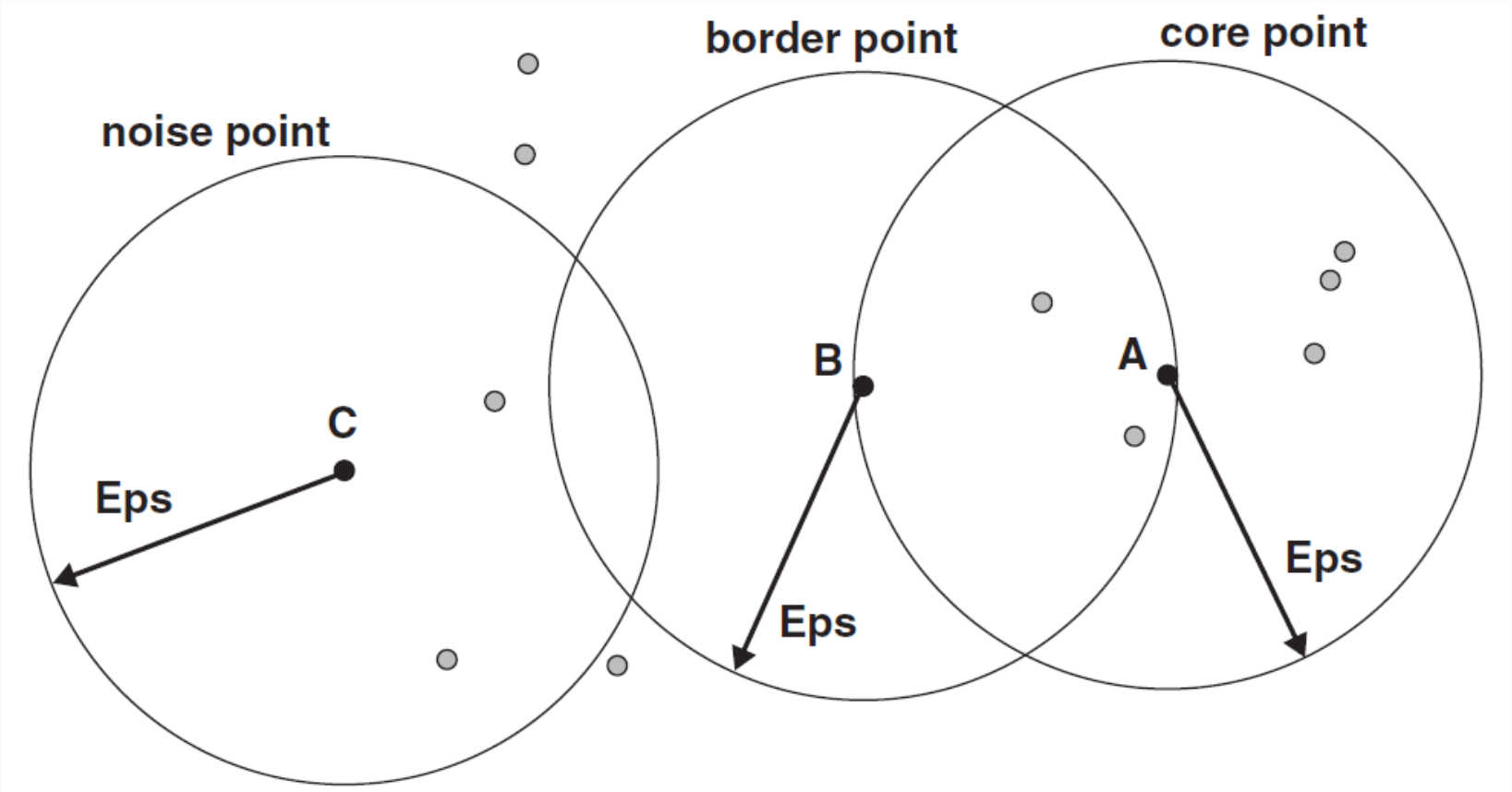

Figura 3: Puntos central, frontera y ruido ($MinPts = 4$)

Clustering basado en densidad

Algorithm 1: DBSCAN

Etiquetar todos los puntos como central, frontera o ruido;
Eliminar los puntos ruido;
Unir los puntos centrales que estén a una distancia menor que *Eps*;
Cada grupo de puntos centrales conectados forma un cluster;
Asignar cada punto frontera al cluster de uno de sus puntos centrales;

Algoritmo DBSCAN. Elección de parámetros (i)

- El comportamiento del algoritmo DBSCAN depende de la elección de sus parámetros (*Eps* y *MinPts*).
- Si estos parámetros no se fijan adecuadamente se obtendrán clusters poco útiles.
- Un procedimiento ampliamente empleado para fijar el valor de estos parámetros usa el concepto de *k*-distancia.
- La *k*-distancia de un punto se define como la distancia al *k*-ésimo punto más cercano.

Algoritmo DBSCAN. Elección de parámetros (ii)

- Tras calcular la k-distancia de cada punto, se ordenan estas de menor a mayor y se muestran en un gráfico.
- El valor en el que se produce un cambio drástico de la curva es un valor apropiado para *Eps*.
- El valor apropiado para *MinPts* es k.

Clustering basado en densidad

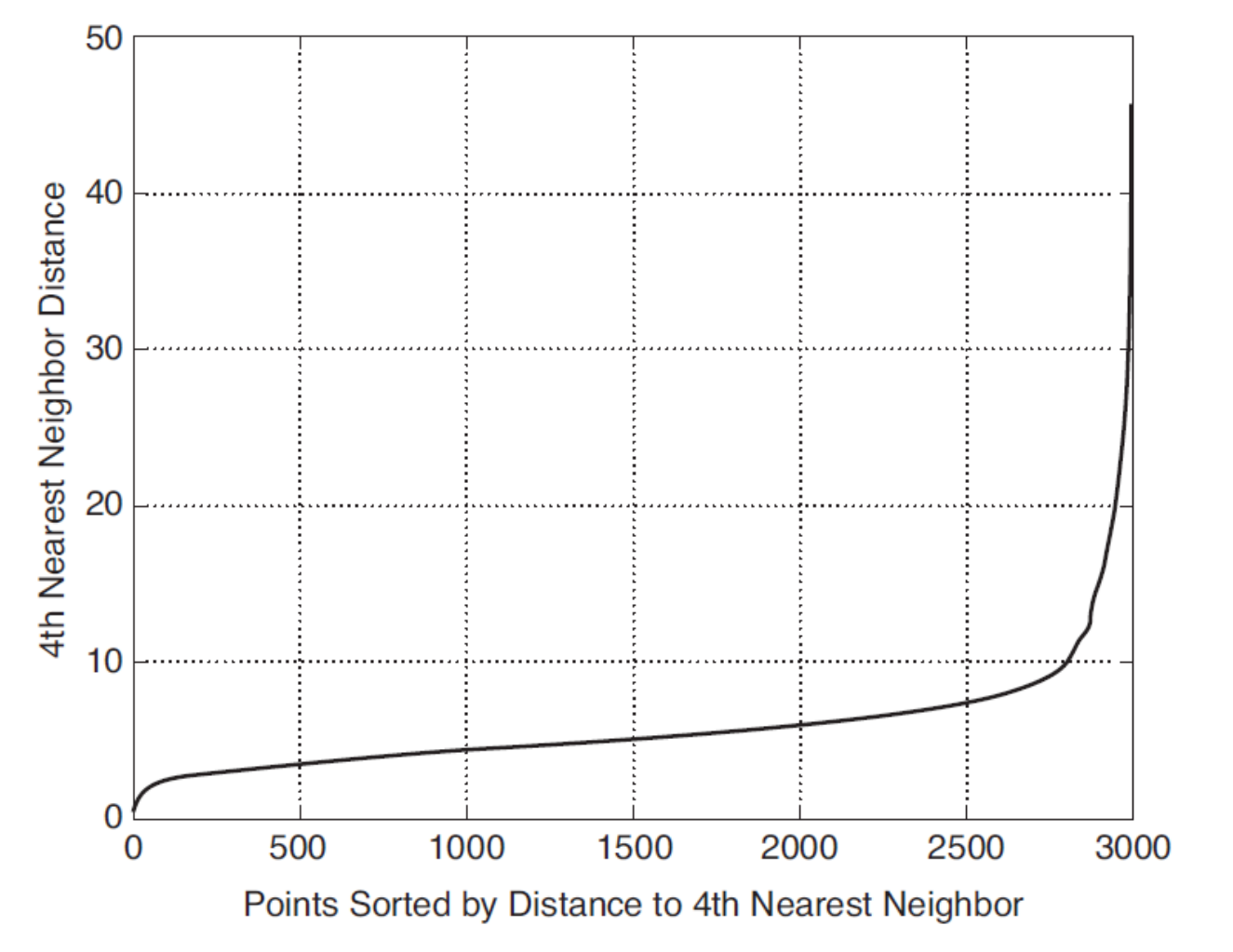

Figura 4: k-distancia ($k = 4$)

BIBLIOGRAFÍA

Bibliografía

Simoudis, E., Han, J., and Fayyad, U. M.
A density-based algorithm for discovering clusters in large spatial databases with noise.
In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (1996), AAAI Press, p. 226231.

Tan, P.-N., Steinbach, M., and Kumar, V.
Introduction to Data Mining.
Addison-Wesley, 2006.

Licencia