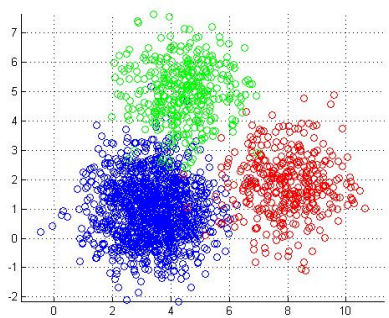
AGRUPACIÓN - ALGORITMOS

Elizabeth León Guzmán

Research Group on Data Mining – MIDAS
Universidad Nacional de Colombia, Bogotá D.C., Colombia

June 4, 2025



Agenda

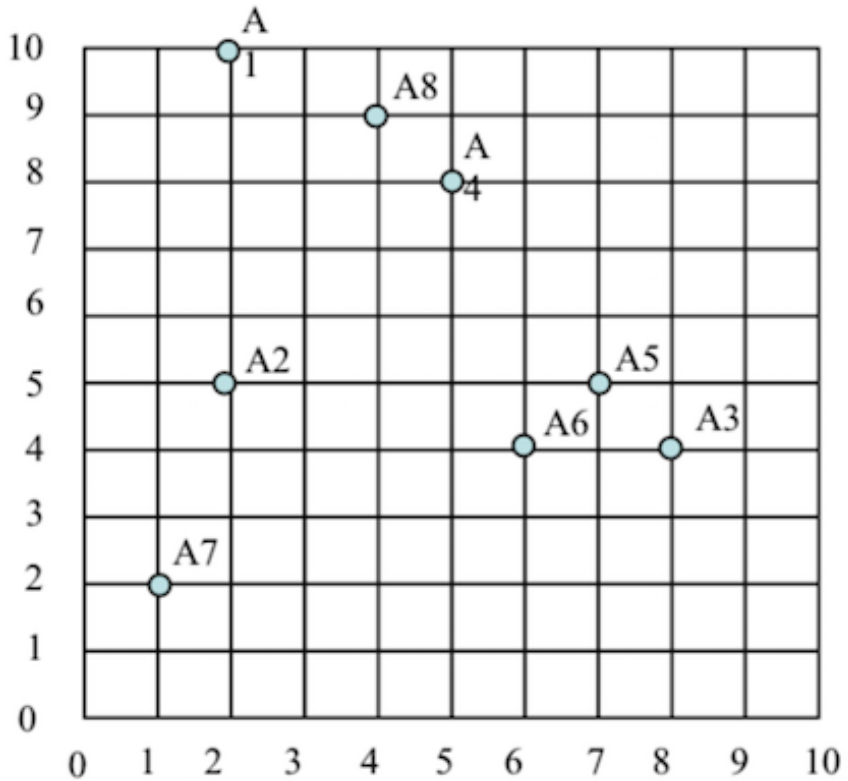
Agenda

Nearest Neighbor clustering

- Calcular la matriz de proximidad
- Seleccionar un punto como primer cluster, k_1
- Seleccionar otro punto j (de acuerdo a un orden) y revisar su distancia con el cluster inicial k_1. Si la distancia es menor del $umbral$, el nuevo punto j es asignado al cluster k_1. En caso contrario el nuevo punto j formará un nuevo cluster,k_2 .
- Seleccionar otro punto (de acuerdo al orden) y asignarlo al cluster más cercano, si la distancia es menor que el $umbral$. En caso del que la distancia sea mayor al $umbral$, el punto formará un nuevo cluster.
- Repetir el paso anterior hasta terminar los puntos.

Nearest Neighbor clustering

Ejemplo: $umbral = 4$

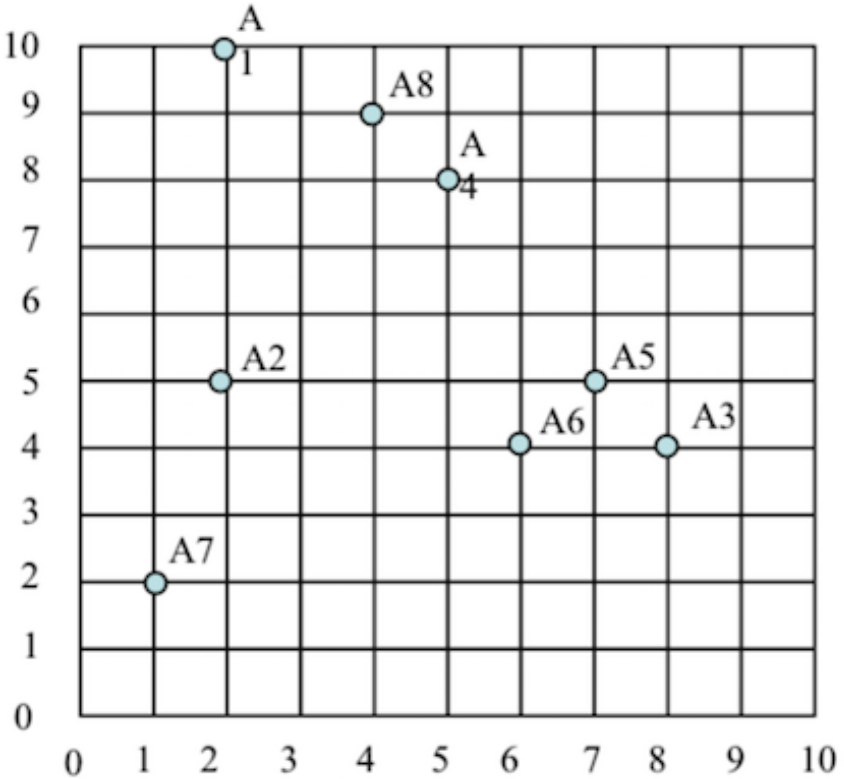


Matriz de proximidad (distancia Euclidean)

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Nearest Neighbor clustering

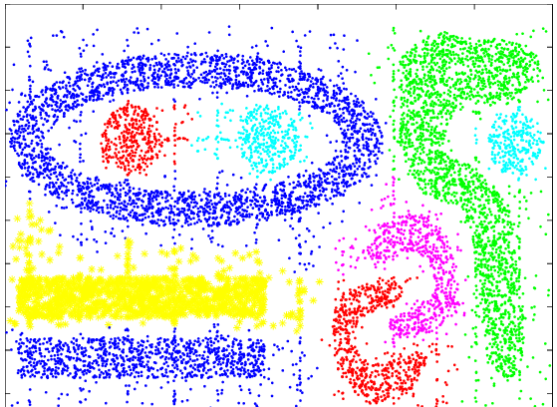
Ejemplo: $umbral = 4$

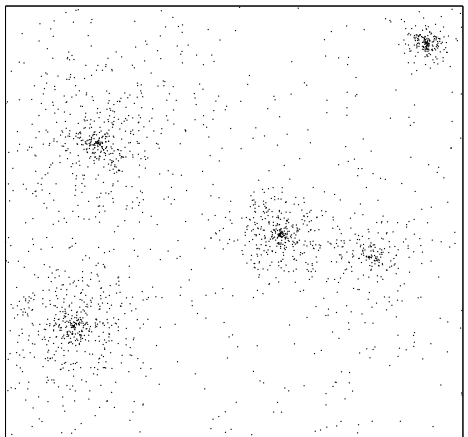


- $K_1 = A1, A8, A4,$
  $K_2 = A2, A7,$
  $K_3 = A3, A5, A6$
- ¡Orden de los datos! ¿Diferentes resultados?
- Clustering On line (streamming)

DBScan

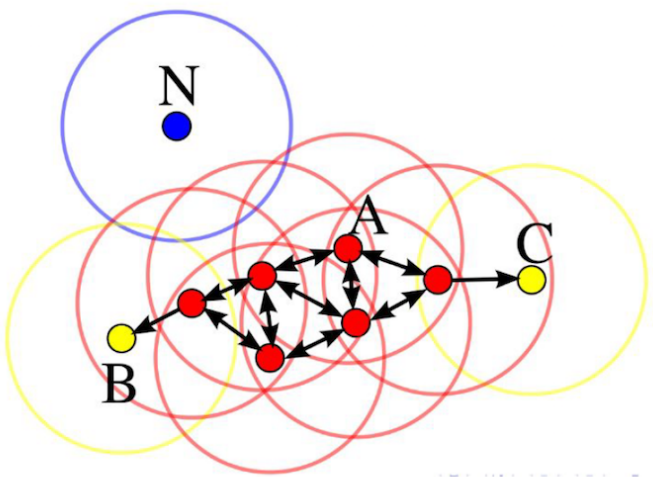
- Algoritmo basado en densidad. Descubrir Clusters en grandes bases de datos espaciales con ruido (DBSCAN)
- Propuesto por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiaowei Xu en 1996.Instituto de Ciencias de la Computación - Universidad de Munich. Ester et al. (1996)
- Los clusters son como cúmulos de alta densidad de puntos. Por lo cual, si un punto pertenece a un clúster, debe estar cerca de un montón de otros puntos de dicho clúster. Kriegel et al. (2011)





DBScan

- Dos parámetros: un número ε positivo y un número natural $minPoints$.
  - ε distancia (radio de un punto seleccionado)
  - $minPoints$ número de puntos mínimo en la distancia ε
- Se elige un punto arbitrario en el conjunto de datos. Si hay una cantidad de puntos mayor o igual a $minPoints$ a una distancia ε del punto arbitrario, a partir de ese momento se consideran todos los puntos como parte de un "cluster".



Tomado de: http://wwwae.ciemat.es/~cardenas/curso_MD/clustering.pdf.

DBScan

- A continuación, se expande ese grupo mediante la comprobación de todos los nuevos puntos y ver si ellos también tienen más puntos minPoints a una distancia épsilon, creciendo el cluster de forma recursiva en caso afirmativo.
- A medida que se generan los clúster, quedan puntos sin añadir al clúster.
- Se elige un nuevo punto arbitrario y se repite el proceso. Es posible que el punto arbitrario escogido tenga menos de $minPoints$ puntos en su círculo de radio ε, y tampoco sea parte de cualquier otra agrupación. Si ese es el caso, se considera un $puntoderuido$ que no pertenecen a ningún grupo.

DBScan

- **Densidad:** Número de puntos en un radio específico (epsilon).
- **Puntos "core":** Puntos interiores de un cluster (cuando tienen, al menos, un número mínimo de puntos $minPoints$ en su vecindario de radio ε,).
- **Puntos "border":** Tienen menos de $minPoints$ puntos en su vecindario de radio ε, estando en el vecindario de algún punto "core".
- **Ruido:** Cualquier punto que no forma parte de un cluster ("core") ni está en su frontera ("border").

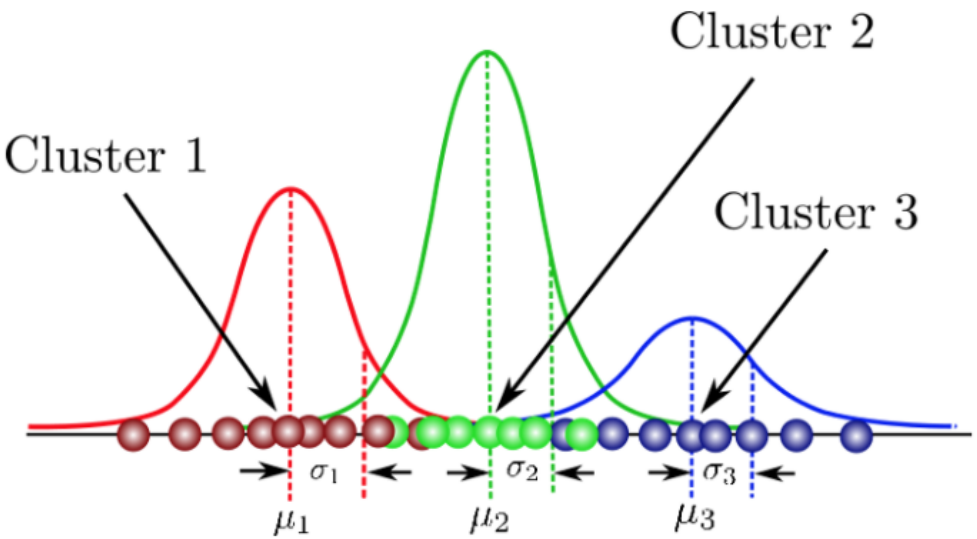
Enlace ejemplo: http://educlust.dbvis.de/#

Gaussian Mixture Model (GMM)

- Asume que los datos provienen de varias subpoblaciones (modeladas separadamente) y toda la población es una mezcla de esas subpoblaciones.
- Cada grupo o cluster corresponde a una subpoblación.
- **Ejemplo:** para agrupación de documentos, GMM encuentra grupos donde un documento puede pertenecer a más de un grupo, y cada documento tiene una representación probabilística a cada uno de los grupos encontrados.

Gaussian Mixture Model (GMM)

- Cada subpoblacion proviene de una distribución **Gaussiana**, y forman un función que son la mezcla de Gaussianas. Cada Gaussiana tiene: media μ, covarianza Σ, y mezcla de probabilidades que indica el tamaño del cluster.



Gaussian Mixture Model (GMM)

El modelo GMM es una generalización de K-means con un enfoque más probabilístico, se asume que las representaciones de los documentos son realizaciones de una variable aleatoria X con distribución $(X|\Theta)$, donde Θ son los parámetros del modelo. La distribución $P(X|\Theta)$ es una función compuesta de las distribuciones de cada cluster $P(X|C_j)$ siguiente forma:

$$P(X = \vec{x}_i|\Theta) = \sum_{j=1}^{k} \lambda_j P(X = \vec{x}_i|C_j)$$

Donde k es el número de clusters y λ_j son constantes de normalización de las distribuciones de cada uno de los clusters C_j.

Gaussian Mixture Model (GMM)

Adicionalmente, para cada una de estas distribuciones se asume normalidad, por ello, cada cluster se representa por medio de una media $\vec{\mu}_j$ y una matríz de covarianza $\mathbf{\Sigma}_j$:

$$P(X = \vec{x}_i | C_j) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_j|}} \exp\left(-\frac{1}{2}(\vec{x}_i - \vec{\mu}_j)\mathbf{\Sigma}_j^{-1}(\vec{x}_i - \vec{\mu}_j)^T\right)$$

Gaussian Mixture Model (GMM)

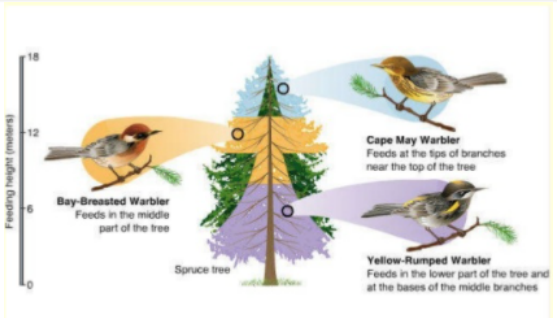
- Finalmente, la estimación de parámetros de un GMM se realiza por medio de algoritmos como *expectation-maximization*, inferencia variacional o estrategias basadas en *Markov-Chain Monte Carlo*.
- GMM es una generalización de K-means, el K-means puede verse como un caso específico de GMM donde las matrices de covarianza $\mathbf{\Sigma}_j$ son matrices identidad (una matriz de ceros con una diagonal de unos), y un caso donde cada punto pertenece únicamente a un único cluster $\vec{x}_i \in C_j \rightarrow P(X = \vec{x}_i | C_j) = 1$.

ECSAGO

Inspirado en la formación de “Nichos” en la naturaleza.

Nicho en la naturaleza

Cada especie pertenece a un niche en una comunidad. Un nicho representa *”un role”* de las especies que incluye tipo de comida, donde viven, donde se reproducen y su relación con otras especies.



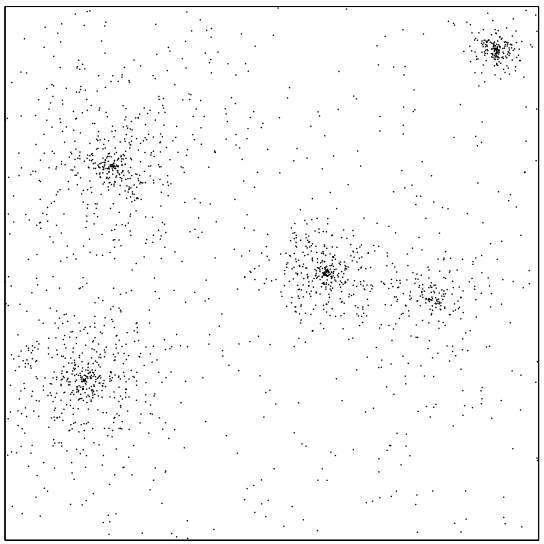
https://socratic.org/questions/what-are-some-examples-of-an-ecological-niche

ECSAGO: ”Genetic Niching”. Leon et al. (2006)

Cada nicho es un grupo (cluster) el algoritmo intenta encontrar los nichos usando un algoritmo evolutivo y una técnica de “niching”

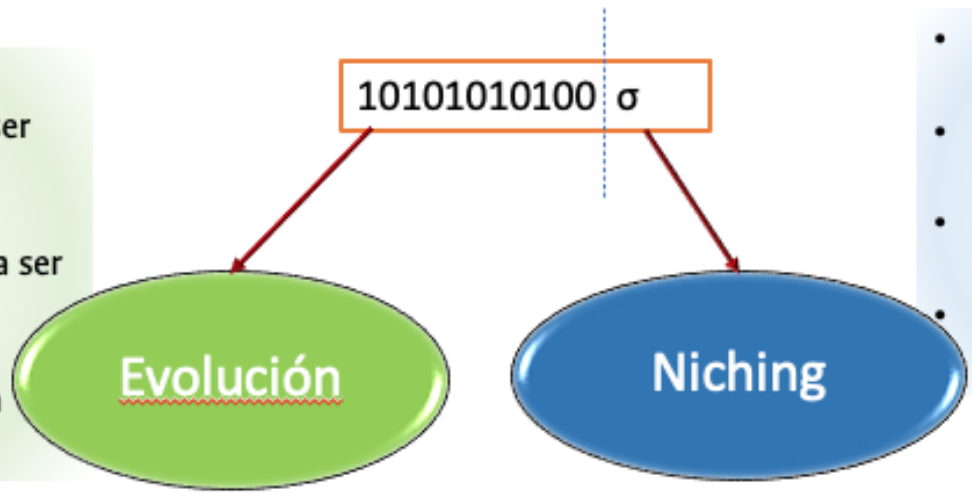
ECSAGO

- "Genetic Niching"
- Basado en:
  - Densidad
  - Conceptos de evolución natural
- Areas densas son clusters
- Clusters circulares $\rightarrow$ centro y radio
- Encuentra centros y radios de los clusters



ECSAGO

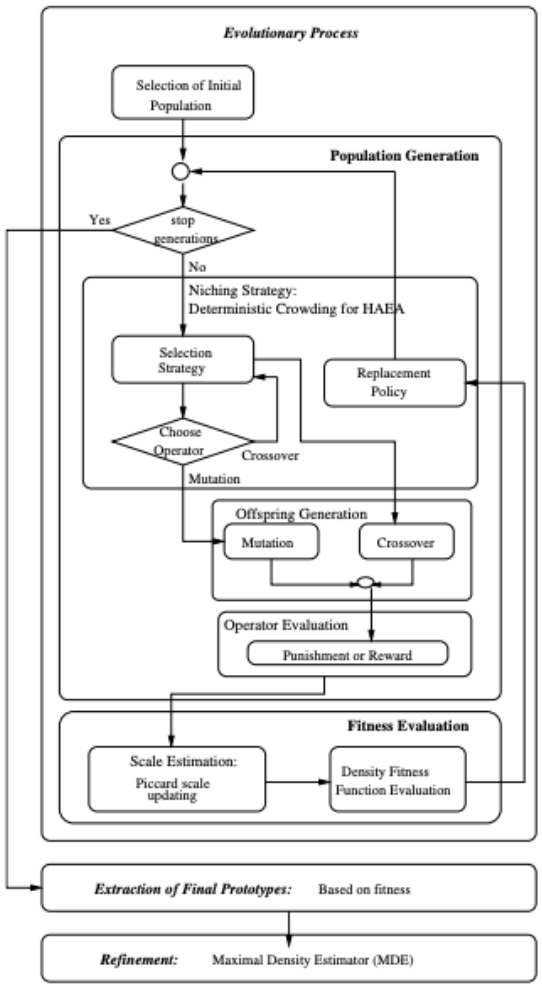
- Encuentra centroides
- **Individuo:** Candidato a ser centroide de un cluster
- **Población:** Conjunto de centroides candidadtos a ser clusters
- **Fitness**: Densidad de un cluster hipotetico en esa localización (centroide)



- Optimización multimodal
- Permite mantener los nichos -> clusters
- Usa Deterministic Crowding
- Hill Climbing para optimizar el radio σ

ECSAGO

Modelo



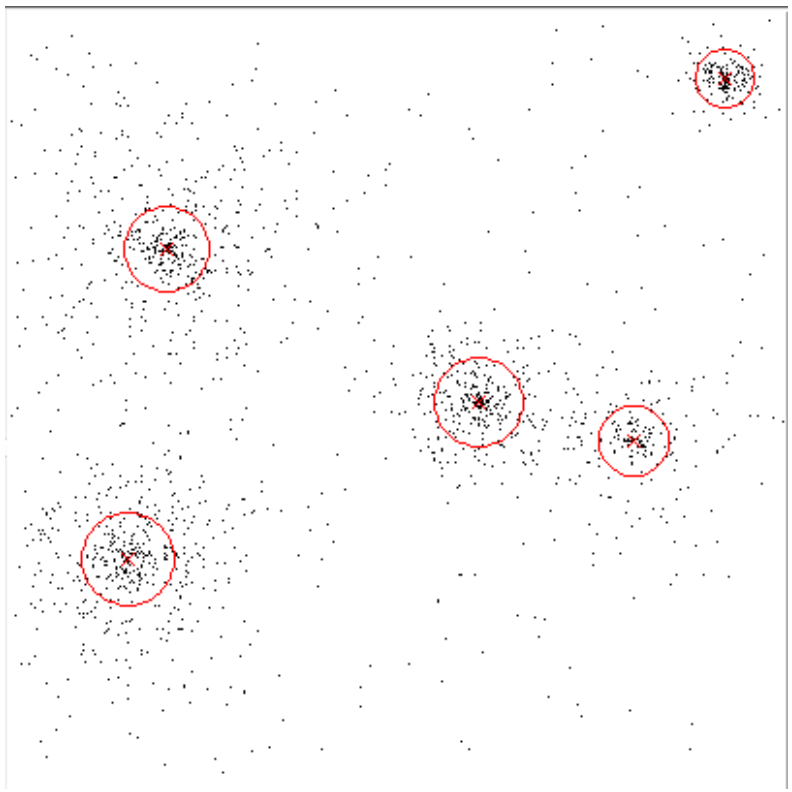
Fitness, pesos y escala (radio σ)

$$f_i = \frac{\sum_{j=1}^{N} w_{ij}}{\sigma_i^2}$$

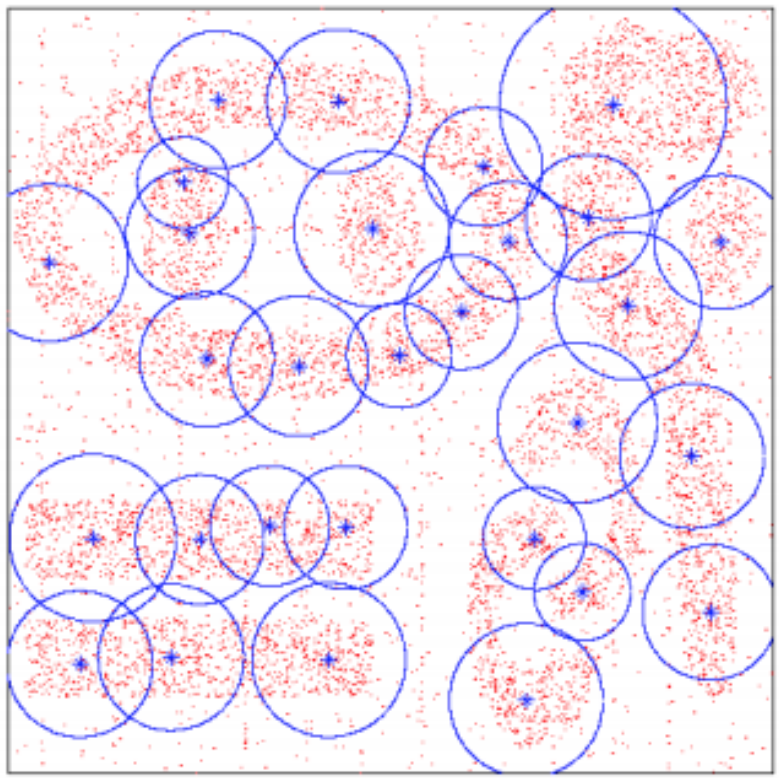
$$w_{ij} = exp(\frac{d_{ij}^2}{2\sigma_i^2})$$

$$\sigma_i^2 = \frac{\sum_{j=1}^{N} w_{ij} d_{ij}^2}{\sum_{j=1}^{N} w_{ij}}$$

ECSAGO - Results



(a) 5 clusters



(b) Chameleon

ECSAGO - Aplicación Agrupación de Documentos

"20 NewsGroup Data set"

Cid	Size	AvgSim	Entpy	Purty	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	comp.windows.x	misc.forsale	rec.autos	rec.motorcycles	rec.sport.baseball	rec.sport.hockey	sci.crypt	sci.electronics	sci.med	sci.space	soc.religion.christian	talk.politics.guns	talk.politics.mideast	talk.politics.misc	talk.religion.misc
0	28	0.249	0.238	0.75	2	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	5
1	40	0.147	0.792	0.225	0	4	3	2	3	3	2	3	0	0	0	0	3	3	1	1	0	0	0	1
2	63	0.133	0.708	0.254	3	1	0	0	0	1	0	0	1	1	0	4	0	7	4	0	11	18	10	4
3	46	0.128	0.913	0.13	1	2	4	4	5	6	0	4	3	2	1	1	2	2	2	3	1	0	1	2
4	62																							
5	25																							
6	52																							
7	29																							
8	20																							
9	38																							
10	35																							
11	45																							
12	35																							

Mayority class
Minority class

Religion.christian

Religion.misc

Politics

>(2) It seems probable that no one displayed the body of Jesus because no
>one kne Paul says to the Corinthians hat "that the gospel will be foolishness to the world, because it is
>explanat spiritually discerned." And so, people without the spirit of God haven't a
>Don't b clue to what the Bible is saying. From your point of view, that's
>gospel, a incredibly circular and convenient. To me, it is mysteriously and supernaturally
>up in or bizarre. I can see it, but you can't. This is not arrogance on
>record t my part. Trust me. It is as bizarre to you as it is to me. But nonetheless,
>guards,
Of course it is a truth, explainable or not.

ECSAGO

Ventajas

- No supervisado en número de clusters. Encuentra el número de clusters automáticamente.
- Robusto al ruido
- Estimación automática del tamaño del niche (cluster)

Limitaciones

- Número de parámetros
- Complejidad

eleonguz@unal.edu.co

www.midas.unal.edu.co

References I

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). 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)* (p. 226-231).

Kriegel, H., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *WIREs Data Mining Knowledge Discovery*, 231-240. doi: 10.1002/widm.30

Leon, E., Nasraoui, O., & Gomez, J. (2006, 01). Ecsago: Evolutionary clustering with self adaptive genetic operators. In (p. 1768 - 1775). doi: 10.1109/CEC.2006.1688521