Anàlisi de clústers: formalisme i aplicacions.

Marc Seguí Coll

Universitat Autònoma de Barcelona

Gener de 2021

Continguts

Què és l'anàlisi de clústers?

Motivació

Si ens fixem en la figura de sota, a simple vista podem distingir quatre grups o clústers de punts.

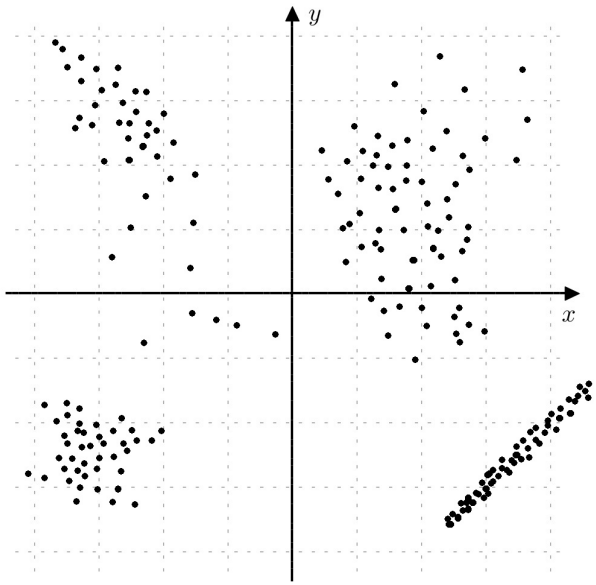

Figura: Mostra bivariada d'un conjunt de dades representada en el pla Cartesià (exemple il·lustratiu).

Què és l'anàlisi de clústers?

Motivació

Si ens fixem en la figura de sota, a simple vista podem distingir quatre grups o clústers de punts.

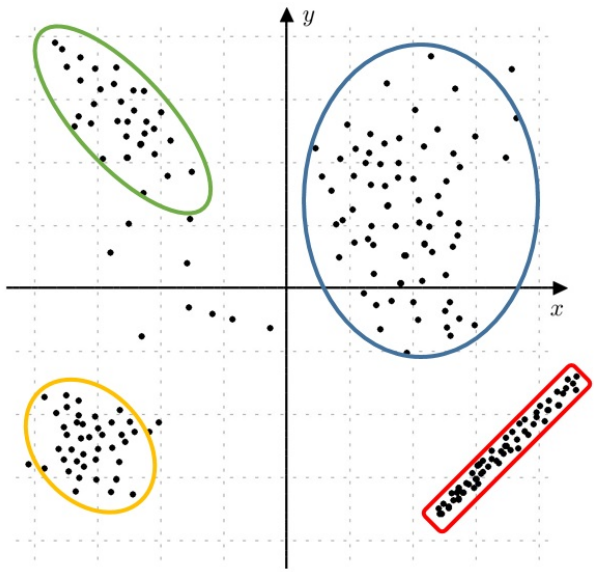

Figura: Mostra bivariada d'un conjunt de dades representada en el pla Cartesià (exemple il·lustratiu).

Temes que aborda l'anàlisi de clústers

Objectius

L'anàlisi de clústers tracta de d'establir criteris i mètodes per tal de classificacar d'elements d'un conjunt de dades en clústers (grups).

Temes que aborda l'anàlisi de clústers

Objectius

L'anàlisi de clústers tracta de d'establir criteris i mètodes per tal de classificacar d'elements d'un conjunt de dades en clústers (grups).

Similitud entre elements

Dos elements pertanyen al mateix clúster si són similars. Intuïtivament pot ser senzill determinar la similitud entre elements. Estudiarem com formalitzar aquests procediments i conceptes.

Temes que aborda l'anàlisi de clústers

Objectius

L'anàlisi de clústers tracta de d'establir criteris i mètodes per tal de classificacar d'elements d'un conjunt de dades en clústers (grups).

Similitud entre elements

Dos elements pertanyen al mateix clúster si són similars. Intuïtivament pot ser senzill determinar la similitud entre elements. Estudiarem com formalitzar aquests procediments i conceptes.

Aplicacions

Tot i que donarem un enfoc més formal d'aquesta branca de l'anàlisi multivariant, avui dia l'anàlisi de clústers és prou emprat en processament d'imatges i *machine learning*.

Exemple d'aplicació

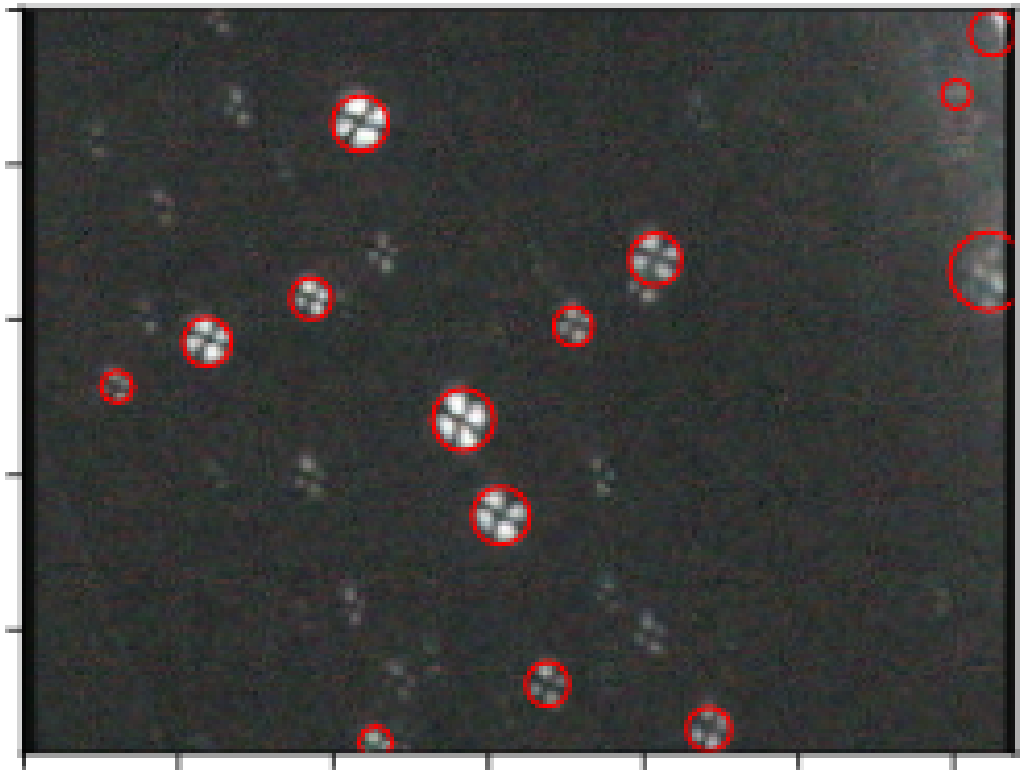

Figura: Selecció de cristalls de tamanys concrets a través d'un algorsime de *clustering* [10].

Definicions

Definició (Patró)

*Un **patró** $\mathbf{x}$ és la representació d'un element η del conjunt d'observacions X. Ens centrarem en el cas en què aquests patrons es poden definir en un espai mètric E de dimensió n, de manera que podem escriure $\mathbf{x} = (x_1, \ldots, x_n) \in E$.*

Definicions

Definició (Patró)

*Un **patró** $\mathbf{x}$ és la representació d'un element η del conjunt d'observacions X. Ens centrarem en el cas en què aquests patrons es poden definir en un espai mètric E de dimensió n, de manera que podem escriure $\mathbf{x} = (x_1, \ldots, x_n) \in E$.*

Definició (Atributs)

*Els **atributs** d'un patró $\mathbf{x}$ són les components (paràmetres) individuals x_i de $\mathbf{x}$. Direm que E és l'**espai d'atributs**.*

Definicions

Definició (Patró)

*Un **patró** $\mathbf{x}$ és la representació d'un element η del conjunt d'observacions X. Ens centrarem en el cas en què aquests patrons es poden definir en un espai mètric E de dimensió n, de manera que podem escriure $\mathbf{x} = (x_1, \ldots, x_n) \in E$.*

Definició (Atributs)

*Els **atributs** d'un patró $\mathbf{x}$ són les components (paràmetres) individuals x_i de $\mathbf{x}$. Direm que E és l'**espai d'atributs**.*

Definició (Mesura de proximitat)

*Una **mesura de proximitat** és una distància d sobre l'espai d'atributs E. Alguns exemples més comuns són la distància Euclidiana o la distància de Mahalanobis.*

Definicions

Observació

Considerem $E = \mathbb{R}^2$ i $E' = \mathbb{R}^+ \times (0, 2\pi)$. Un punt bidimensional η es pot representar $\mathbf{x} = (x, y) \in E$ ò $\mathbf{x} = (r, \theta) \in E'$. Tot i això no té perquè satisfer-se la igualtat $d(x, y) = d(r, \theta)$.

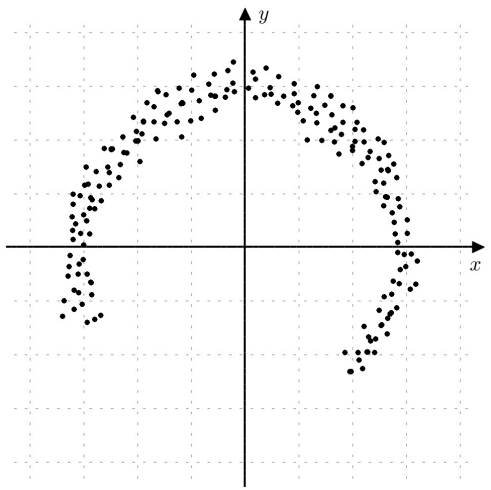

Figura: Dades corresponents a una mostra bivariant representada al pla Cartesià. Les dades estansituades concèntricament (exemple il·lustratiu).

Algorismes i mètodes

Descripció dels mètodes de *clustering*

Sigui $\mathfrak{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\} \subset E$ el conjunt de patrons del nostre experiment amb $\mathbf{x_j} = (x_{j_1}, \ldots, x_{j_n})$ per a tot $j \in \{1, \ldots, m\}$. Busquem trobar una partició en clústers $\mathfrak{C} = \{C_1, \ldots, C_p\}$ de $\mathfrak{X}$ que compleixi els criteris de similitud establerts. Cada algorisme empra un criteri de selecció per decidir si dos elements són similars o no.

Algorismes i mètodes

Descripció dels mètodes de *clustering*

Sigui $\mathfrak{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\} \subset E$ el conjunt de patrons del nostre experiment amb $\mathbf{x_j} = (x_{j_1}, \ldots, x_{j_n})$ per a tot $j \in \{1, \ldots, m\}$. Busquem trobar una partició en clústers $\mathfrak{C} = \{C_1, \ldots, C_p\}$ de $\mathfrak{X}$ que compleixi els criteris de similitud establerts. Cada algorisme empra un criteri de selecció per decidir si dos elements són similars o no.

Tipus d'algorismes

Generalment, es distingeixen dos formes d'atacar el problema: per mitjà d'algorismes jeràrquics o a través d'algorismes particionals. Tot i això, recentment, s'han proposat algorismes basats en la densitat o en la distribució dels patrons.

Descripció dels algorismes

Algorismes jeràrquics

Aquest tipus d'algorismes poden ser o bé **aglomeratius** o bé **divisius**. Sense entrar gaire en detall, a cada pas de l'algorisme s'agrupen o es separen clústers segons la distància entre aquests. Per poder fer això necessitem d'una **distància entre clústers** que dependrà de l'algorisme emprat.

Descripció dels algorismes

Algorismes jeràrquics

Aquest tipus d'algorismes poden ser o bé **aglomeratius** o bé **divisius**. Sense entrar gaire en detall, a cada pas de l'algorisme s'agrupen o es separen clústers segons la distància entre aquests. Per poder fer això necessitem d'una **distància entre clústers** que dependrà de l'algorisme emprat.

Algorismes particionals

Parteixen d'un nombre K prefixat de clústers i tracten de determinar una única partició on els clústers siguin el més similars possible.

Exemples d'algorismes jeràrquics

- ***Clustering* per enllaç simple.** La distància entre dos clústers A i B és la mínima distància entre patrons de cada clúster.

Exemples d'algorismes jeràrquics

- ***Clustering* per enllaç simple.** La distància entre dos clústers A i B és la mínima distància entre patrons de cada clúster.
- ***Clustering* per enllaç complet.** En aquest cas, en lloc de la mínima distància, s'empra la màxima (patrons allunayts).

Exemples d'algorismes jeràrquics

- ***Clustering* per enllaç simple.** La distància entre dos clústers A i B és la mínima distància entre patrons de cada clúster.
- ***Clustering* per enllaç complet.** En aquest cas, en lloc de la mínima distància, s'empra la màxima (patrons allunayts).
- ***Clustering* per enllaç promig.** Es promitgen les distàncies entre cada parella de patrons formada per un patró de cada clúster.

Exemples d'algorismes jeràrquics

- ***Clustering* per enllaç simple.** La distància entre dos clústers A i B és la mínima distància entre patrons de cada clúster.
- ***Clustering* per enllaç complet.** En aquest cas, en lloc de la mínima distància, s'empra la màxima (patrons allunayts).
- ***Clustering* per enllaç promig.** Es promitgen les distàncies entre cada parella de patrons formada per un patró de cada clúster.

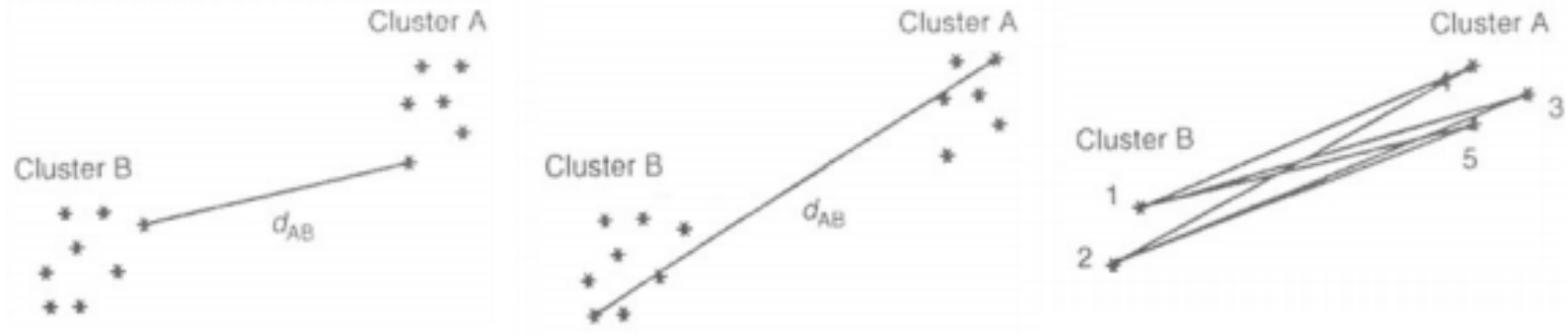

Figura: Distàncies segons: enllaç simple, enllaç complet, enllaç promig [3].

Algorisme particionals K-*Means*

Definim l'**error quadràtic** de la partició $\mathfrak{C} = \{C_1, \ldots, C_K\}$ com

$$\mathrm{SE}\left(\mathfrak{C}\right) = \sum_{j=1}^{p} \sum_{\mathbf{x} \in C_j} d\left(\mathbf{x}, \mathbf{c_j}\right)^2$$

amb $\mathbf{c_j}$ el centroide del clúster C_j. El centroide és el centre de gravetat del clúster. Formalment, si el clúster C_j conté n_j patrons:

$$\mathbf{c_j} = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

Algorisme K-means. Tria aleatòriament (o es proporciona) una partició inicial en K-clústers. S'hi realitzen modificacions fins que SE s'estabilitza o els centroides entre pas i pas no canvien.

Mètode K-*Means* per determinar exoplanetes habitables

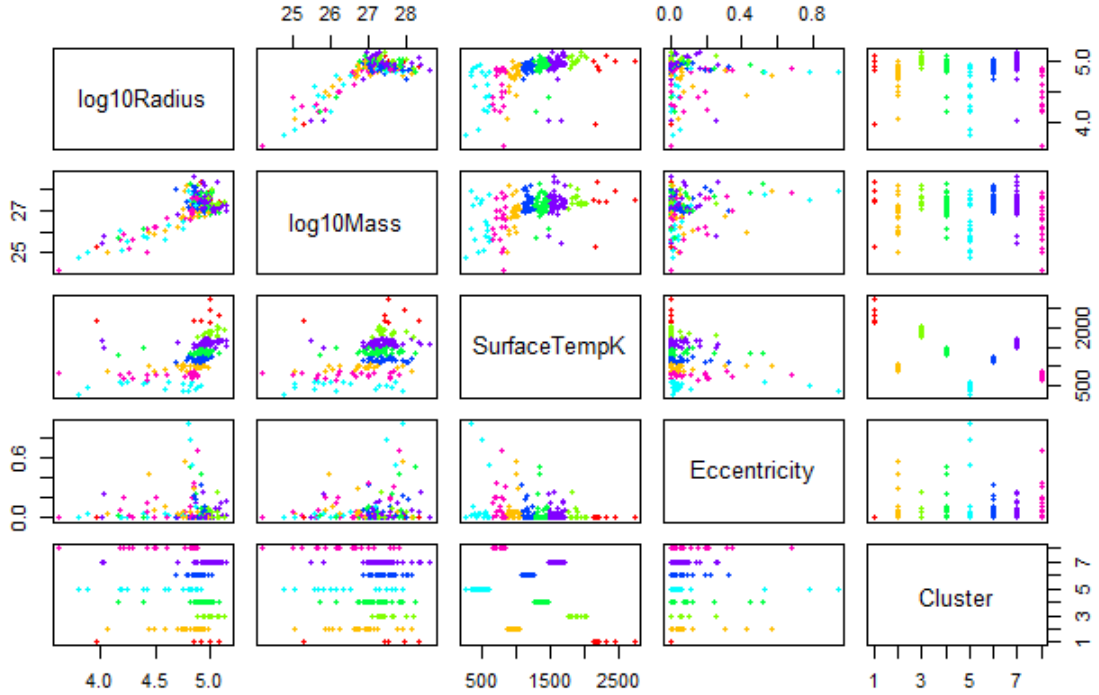

Figura: A partir d'un conjunt de dades [9] hem determinat 17 planetes habitables mitjançant K-*Means* amb $K = 8$. Hem considerat que un planeta és habitable si els atributs **radi**, **massa**, **temperatura de la superfície** i **excentricitat** són similars als de la Terra.

Bibliografia i referències

[1] A. K. Jain, R. C. Dubes. Michigan State University. *Algorithms for Clustering Data*. Prentice Hall (1988)

[2] W. K. Härdle, L. Simar. Humboldt-Universität zu Berlin, Katholieke Universiteit Leuven. *Applied Multivariate Statistical Analysis*. Springer (2012)

[3] B. S. Everitt. King's College. *An R and S-PLUS® Companion to Multivariate Analysis*. Springer (2005)

[4] A. K. Jain, M. N. Murty, P. J. Flynn. Michigan State University, Indian Institue of Science, The Ohio State University. *Data Clustering: A Review*. ACM Computing Surveys (1999)

[5] D. L. Davies, D. W. Bouldin. University of Tennessee. *A Cluster Separation Measure*. IEEE (1979)

Bibliografia i referències

[6] A. Fernández, S. Gómez. Universitat Rovira i Virgili. *Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms*. Journal of Classification (2008)

[7] H. Kriegel, P. Kröger, J. Sander, A. Simek. Ludwig-Maximilians-Universität München, University of Alberta. *Density-based clustering*. WIREs Data Knowledge Discovery (2011)

[8] P. Dyches, F. Chou. Jet Propulsion Laboratory, NASA Headquarters. *The Solar System and Beyond is Awash in Water* NASA (2015)

[9] A. Tribick, C. Sturm, H. Rein and more. Múltiples organitzacions. *Open Exoplanet Catalogue Database: All extrasolar planets* Open Exoplanet Catalogue (2021)

Bibliografia i referències

[10] C. Falcó, M. Seguí, S. Serrano. Universitat Autònoma de Barcelona. *Estats metaestables: obtenció dels estats amorf i cristal·lí en un polímer* Laboratori de Termodinàmica de la UAB (2019)