

---

**Abstract**

Clusterización (qué es y para qué sirve) - Técnicas (para qué sirven) - Paquetes

*Keywords:* Data Mining, Clustering, R

---

**1. Introducción**

Desde finales del siglo XX se ha considerado que vivimos en la “era de la información”, una etapa caracterizada por el incremento, desarrollo y propagación de emergentes tecnologías de la información y comunicación que han permitido al ser humano romper las barreras de la distancia, el tiempo y el lugar a la hora de comunicarse y compartir información; actividades que han sido decisivas en nuestra historia [1]. Sin embargo, la era en la que realmente vivimos es la “era de los datos”, donde cada día se generan más de dos mil quinientos millones de petabytes<sup>1</sup> de datos provenientes de comercios, ciencias, Internet y casi cualquier actividad del día a día [2] que acaban volcados en redes de ordenadores, sitios web, bases de datos y otros medios de almacenaje.

Esta explosión de datos, a la que se ha denominado Big Data, se debe al alto grado de computarización de la sociedad y el avance de herramientas de recolección y almacenamiento de datos. Negocios en todo el mundo generan grandes cantidades de datos derivados de transacciones, stock de productos, platillas de empleados, etc. Las ramas de la ciencia producen datos de manera constante frutos de experimentos, observaciones, recogida de muestras, etc. Y más recientemente, Internet y las redes

sociales han sido las principales responsables del aumento excesivo de datos, siendo usadas por millones de personas simultáneamente.

Y, aunque esto ha supuesto una considerable mejora para la humanidad pues la información nunca había sido tan accesible, también ha traído consecuencias negativas y problemas como el almacenamiento y organización de los datos, datos no estructurados que entorpecen su acceso y procesamiento, dificultades a la hora de analizar los datos apropiadamente pudiendo generar desinformación y complicaciones para mostrar los resultados de forma apropiada y aplicarlos de manera eficiente y útil en el mundo real [3].

Como resultado, ha surgido una nueva ciencia que se ha posicionado rápidamente como una de las disciplinas más influyentes de la actualidad: Data Science (Ciencia de los Datos), que debido a su reciente aparición, carece de una definición consensuada, pero podríamos concretarla como “*Ciencia que usa Estadística, Inteligencia Artificial, Programación y Bases de Datos para posibilitar la extracción de conocimiento a partir de datos*” [4]. A su vez, dentro de esta ciencia se han desarrollado otras tres ramas: Data Warehousing, Data Mining y Visualization; cada una de ellas enfocada a resolver o afrontar uno o varios de los problemas mencionados previamente: organización y agrupación de datos, análisis de los mismos y presentación de los resultados, respectivamente.

De entre estas nuevas disciplinas, Data Mining es la que se centra en el procesamiento de los datos, procedimiento por el cual se obtiene la información, y se podría definir como “*Proceso de descubrimiento de patrones interesantes y conocimiento a*

---

*Email addresses:* aaron.casado@uah.es (Aarón Casado Monge), jcg@uah.es (Juan José Cuadrado Gallego)

<sup>1</sup>Un Petabyte es una unidad de información o almacenamiento de datos equivalente a un cuadrillon de bytes, mil terabytes o un millón de gigabytes. En este caso, es el equivalente a 2.5 quintillones de bytes.

partir de grandes volúmenes de datos” [5], donde dentro de la misma podemos encontrar diferentes técnicas para encontrar patrones y relaciones, y dependiendo de cuál se aplique se puede obtener un resultado totalmente diferente incluso con el mismo conjunto de datos, por lo que es fundamental emplear la técnica apropiada en función del del objetivo a conseguir, los datos con los que se pretende trabajar y el ámbito de aplicación.

Y en algunos de los campos más importantes del mundo moderno como la analítica de negocios, el reconocimiento de imágenes, las búsquedas web, seguridad, biología y ciencias de la salud; existen dificultades a la hora de clasificar o agrupar ciertos datos porque estos no disponen de una etiqueta o valor conocido por el que se pueda hacerlo, pues este no existe o no ha sido definido. Para poder afrontar este problema se utiliza la técnica de Clustering, que permite exactamente generar valores y etiquetas para un conjunto de datos realizando agrupaciones denominadas *clusters* o grupos, donde los datos de un mismo cluster sean muy similares entre ellos y a su vez tengan diferencias claras con datos de otros clusters, permitiendo que cada grupo resultante puede ser etiquetado y tratado como una clase propia.

De esta manera, el objetivo de este Trabajo de Fin de Grado (TFG) es realizar un estudio sobre Clustering, exponiendo de manera teórica qué es este método y qué tipo de utilidades tiene, así como el desarrollo de las diferentes técnicas que existen y las aplicaciones que estas ofrecen. Posteriormente se explorará este método dentro de Bioinformática, un campo centrado en desarrollar técnicas y programas software para analizar datos biológicos, viendo qué aporta a dicha disciplina y cuáles de las técnicas expuestas previamente se emplean y por qué.

Una vez realizado el marco teórico previo, se pretende hacer una parte práctica los diferentes paquetes que ofrecen técnicas de Clustering tanto de carácter general como dentro de Bioinformática para el lenguaje de programación R [6], uno de los más relevantes dentro de Data Science.

## 2. Clustering

Clustering o Cluster Analysis, en español Agrupamiento o Análisis de Grupos y adaptado al término Clusterización, es un método de Data Mining que se basa en el Aprendizaje Automático (Machine Learning), una rama de la Inteligencia Artificial (Artificial Intelligence) que pretende desarrol-

lar sistemas que resuelvan problemas basándose en los resultados de experiencias previas, aprendiendo de sus errores. Concretamente, se fundamenta en uno de los métodos de aprendizaje explorados en esta disciplina: Aprendizaje No Supervisado, enfocado en el desarrollo y descubrimiento de nuevos conocimientos. Este, a diferencia de otros métodos de aprendizaje, no dispone de conocimiento previo sobre lo que poder aprender, por lo que su objetivo principal es discernir patrones y relaciones entre los datos para poder separarlos [7].

En esencia, el proceso de Clustering es el mismo, hasta el punto en que básicamente podríamos considerarlos sinónimos, puesto que Clusterización se aplica principalmente sobre conjuntos de datos que se desean organizar en diferentes grupos pero los valores que delimitan cada cluster son desconocidos y por lo tanto se definen en el propio proceso de clasificación.

De una manera más técnica, podríamos decir que Clustering busca definir para una determinada característica<sup>2</sup> o Suceso Elemental<sup>3</sup> (SE), un conjunto de grupos de observaciones (suceso)<sup>4</sup> con valores cercanos; donde los clusters permiten, dados los diferentes sucesos elementales que configuran un suceso, asignar cada SE al mismo cluster [8].

Para poder decidir si dos datos deben ser agrupados en el mismo cluster o por el contrario, separados, es necesario introducir los conceptos de similitud y disparidad; cuanto más similares sean los datos, más tendrán en común y por lo tanto es más probable que terminen bajo el mismo cluster, y por otra parte, si los datos son diferentes entre sí, serán separados en clusters diferentes. Este criterio se basa en la proximidad de dichos objetos. Si midiéramos la similitud entre dos objetos  $i$  y  $j$ , esta devolvería 0 si ambos fueran totalmente distintos, y cuanto más elevado fuera el valor, más semejantes serían, siendo 1 el valor más alto, indicando que ambos datos son idénticos. De la misma manera, medir la disparidad entre los objetos daría resultados opuestos, con un 0 indicando que son idénticos y con un 1 que no tienen nada en común.

Esta forma de calcular la proximidad no solo

<sup>2</sup>Dicho de una cualidad que da carácter o sirve para distinguir a algo o alguien de sus semejante.

<sup>3</sup>Cada uno de los resultados más simples que se pueen obtener de la realización de un experimento. Obtener el número 3 al tirar un dado.

<sup>4</sup>Se denomina así al subconjunto total de resultados posibles al realizar un experimento. Obtener un 3 o sacar par al tirar un dado.

ayuda a la hora de agrupar los datos en clusters y formar subconjuntos a partir de los resultados, sino que con este proceso también somos capaces de detectar outliers, datos anómalos que pueden ser datos erróneos procedentes de errores de medida o fallos que deben ser eliminados o, datos correctos sumamente importantes que son diferentes al esto y deben ser analizados detenidamente. A la hora de agrupar los datos, puede que queden varios objetos sueltos sin pertenecer a ningún cluster, esos son los datos anómalos.

Además, Clustering también sirve como primera aproximación a la hora de procesar los datos, porque aunque de por sí, este proceso puede aportar información útil agrupando datos similares y clasificándolos, permitiendo un análisis de los resultados que puede dar lugar al descubrimiento de nuevos conocimientos, también se usa como método de preprocesamiento de datos para otros algoritmos de Data Mining que trabajarán sobre las clusters generados y los atributos seleccionados como criterio a la hora de crearlos, pues estos se pueden considerar clases nuevas de objetos.

Es gracias a estas características que Clusterización es empleada en muchos ámbitos del mundo actual, siendo la técnica de Aprendizaje No Supervisado más extendida. Campos como la analítica de negocios, el reconocimiento de imágenes, las búsquedas web, seguridad, biología y ciencias de la salud hacen uso habitual de esta técnica para clasificar tipos de consumidores que comparten preferencias, aunar bajo un mismo subconjunto muchas formas diferentes de escribir el mismo carácter para facilitar el reconocimiento de textos escritos a mano, agrupar resultados similares de una consulta en internet y mostrar los más relevantes dentro de cada grupo o el estudio de la taxonomía<sup>5</sup> de las especies. También se puede hacer uso de su capacidad para detectar datos anómalos con la finalidad de revelar posibles fraudes o transacciones financieras sospechosas, incluso ayudar en la disminución de crímenes analizando los resultados obtenidos tras clusterizar datos de detenciones y delitos; incluso aplicada a datos geofísicos<sup>6</sup> para interpretarlos y obtener resultados significativos. Asimismo, se utiliza a la hora de comparar comunidades en redes

<sup>5</sup>Ciencia que trata de los principios, métodos y fines de la clasificación. Se aplica en particular, dentro de la biología, para la ordenación jerarquizada y sistemática, con sus nombres, de los grupos de animales y de vegetales.

<sup>6</sup>La Geofísica es la ciencia que estudia la Tierra desde el punto de vista de la física.

sociales permitiendo recomendar a los usuarios contenido de su agrado, o dentro del mundo sanitario, ayudando en la identificación y control de diversos tipos de enfermedades e incluso puede usarse como apoyo en la gestión de edificios públicos como bibliotecas, agrupando a los lectores por sus preferencias de manera similar a los consumidores de un negocio [9–13].

Comprimiendo toda esta información, podemos concretar los tres principales motivos por los que Clustering es empleado: clasificación natural tanto de individuos, como de objetos, etc.;, compresión de información para mayor facilidad a la hora de mostrar resultados y generación de una estructura base para el posterior tratamiento de los datos [14].

La utilidad de este método y la flexibilidad que ofrece con las diversas técnicas y formas de aplicarlo de las que dispone e es también parte fundamental de que sea tan comúnmente utilizado y con objetivos tan dispares en gran parte de las áreas del conocimiento. Sin embargo, también es un método frágil, pues depende en gran medida de los datos con los que se trabaje y el criterio escogido para determinar si dos objetos deben agruparse bajo el mismo cluster o separarlos, Por lo que es necesario seguir una serie de pasos a la hora de realizar un Análisis de Grupos de manera correcta [15]:

1. Primero, hay que seleccionar cuidadosamente los datos con los que se va a trabajar, puesto que tanto el tipo de dato como la cantidad de los mismos influyen directamente en los resultados. Si se utilizan demasiados datos el procesamiento puede tener un coste computacional alto y es difícil ofrecer una visualización elegante de los resultados. Pero si se escogen pocos datos se pierde información útil y puede dar lugar a equivocaciones. Es por ello que este paso suele realizarlo un experto en el sector o con su ayuda.
2. Segundo, se debe elegir la forma en la que se va a calcular la similaridad entre los diferentes datos, que analizaremos más adelante cuando analicemos las diferentes técnicas de Clustering.
3. Tercero, debe definirse el criterio con el que se va a tomar la decisión de agrupar en el mismo cluster dos datos. Es decir, escoger el umbral de similaridad a partir del cual dos datos pasan a formar parte del mismo cluster y qué característica o conjunto de ellas van a ser utilizadas para medir la similitud. La elección del umbral

suele ser complicada, y es necesaria mucha experiencia o varias iteraciones de ensayo y error para encontrar un valor correcto.

4. Cuarto, hay que optar por uno de los diversos algoritmos de Clusterización que existen, pues los resultados pueden cambiar considerablemente al variar la estrategia con la que se realiza la agrupación. Estos se verán a continuación.

5. Por último, una vez obtenidos los resultados hay que validarlos e interpretarlos. La resolución de este paso depende del objetivo inicial por el que se haya decidido realizar el análisis.

Como puede verse, tanto el paso inicial como el último requieren de la ayuda de personas cualificadas si pretenden realizarse correctamente, y es en los pasos intermedios donde la intervención de computadoras y programas informáticos son más útiles y están más desarrollados, pues permiten aligerar considerablemente la carga de trabajo y como consecuencia, se puede analizar una mayor cantidad de datos. Pero no deja de ser una tarea complicada que ha ocasionado una gran demanda de expertos en Data Mining y Data Science en diversas áreas de trabajo durante los últimos años.

La utilidad que aporta Clustering permitiendo crear grupos con datos similares entre sí y dispares con respecto a otros grupos con el fin de clasificarlos y las ventajas que esto presenta queda reflejada en la cantidad de usos que recibe, por lo que vamos a profundizar en las diferentes técnicas que son empleadas para la generación de los clusters y su funcionamiento que lo hacen tan importante.

### 3. Técnicas de Clustering

Existe una gran variedad de técnicas de Clustering que cumplen con el propósito de clasificar un conjunto de datos, pero no todas son iguales. Existen una serie de requisitos típicos que debe cumplir un algoritmo de Clusterización para satisfacer las expectativas y realizar un buen análisis. Sin embargo, cumplir todos estos requerimientos de forma simultánea es una tarea casi imposible, por lo que analizar cuáles ofrece cada algoritmo nos ayuda a identificar sus puntos fuertes y débiles y por ende, facilitar la elección del algoritmo apropiado para cada ocasión. Veamos cuáles son dichos requisitos y cómo podemos comparar algoritmos.

#### 3.1. Requisitos de un algoritmo de Clustering

[16] Sobre el papel, analizar datos no debería suponer un problema; pero la realidad es que existen trabas que complican este procedimiento tales como la cantidad de datos a analizar o el tipo de datos. Algunos de los algoritmos de Clustering iniciales no tenían en cuenta este tipo de problemas, porque no existían en su momento, pero ahora es necesario afrontarlos para que el análisis de los datos sea eficiente y produzca resultados completos y veraces.

A continuación, se exponen los requisitos típicos que se exigen dentro de Data Mining a un algoritmo de Clustering, así como debilidades que presentan algunos de los existentes.

- **Escalabilidad:** Muchos algoritmos de Clusterización trabajan bien con conjuntos de datos pequeños que contienen solo unos cuantos cientos de objetos; sin embargo, la gran mayoría de bases de datos contienen millones y miles de millones de datos. Analizar solo un pequeño porcentaje de esos datos puede resultar en conocimiento parcial, siendo necesario que sean escalables para resultados óptimos.
- **Habilidad para lidiar con diferentes tipos de atributos<sup>7</sup>:** Coloquialmente se suele entender por atributo un dato<sup>8</sup> cuantitativo, es decir, un número. Sin embargo, existen otros tipos de datos: nominales<sup>9</sup>, binarios<sup>10</sup>, ordinales<sup>11</sup>, imágenes, gráficos, documentos y mezclas de varios de estos tipos. Por lo que un algoritmo de Clustering debe estar preparado para lidiar con todos o la gran mayoría de ellos.
- **Descubrimiento de clusters con formas arbitrarias:** Gran parte de los algoritmos utilizan formas de calcular la similaridad entre

<sup>7</sup>Dicho de una característica cualitativa de un individuo o dato usada para distinguirlo de una variable o característica cuantitativa.

<sup>8</sup>Valor obtenido para una característica en una observación.

<sup>9</sup>Dicho de un dato cualitativo que proporciona suficiente información para nombrar a una característica y poder diferenciarla de otra.

<sup>10</sup>Datos que solo pueden tomar los valores cero y uno (0, 1).

<sup>11</sup>Datos cualitativos que proporcionan suficiente información para ordenar las observaciones.

los datos basadas en la distancia entre los mismos, lo que suele dar lugar a clusters con formas redondas o esféricas. En determinadas ocasiones esto puede resultar contraproducente y erróneo, por lo que también es preciso que los algoritmos sean capaces de identificar diferentes formas de clusters, no solamente formas circulares o geométricas.

- **Requirir información al usuario:** En algunos algoritmos de Clustering es común pedir cierto tipo de información al usuario a la hora de trabajar con los datos, como el número de clusters que se quieren generar. Este tipo de práctica puede ocasionar problemas dada la complejidad de esta decisión, sobre todo en escenarios en los que los datos cuentan con múltiples atributos. Por lo que es recomendable que los algoritmos eviten este comportamiento tanto para aliviar a los usuarios como para no comprometer la calidad a la hora de formar clusters, pues el algoritmo se adaptaría al valor introducido por el usuario, aunque fuese erróneo, generando también, un resultado errado.

- **Capacidad para trabajar con ruido en los datos:** Aunque es común tratar de antemano los datos para eliminar cualquier tipo de dato no deseado, recordemos que esta técnica también se aplica como método de preprocesamiento para otros algoritmos de Data Mining, por lo que es necesario que no sean sensibles hacia estas impurezas, puesto que podría resultar en una clasificación pobre.

- **Clusterización incremental e insensibilidad al orden de entrada:** Cuando se aplica Clustering en entornos en constante movimiento como el de los negocios, es común que se introduzcan datos nuevos junto a los ya existentes. Ciertos algoritmos no son compatibles con este incremento de información, y es necesario volver a iniciar un nuevo proceso de Clusterización. Mientras que otros que sí toleran este cambio, resulta que son sensibles respecto al orden con el que se han introducido los datos y pueden generar agrupaciones erróneas. Estos problemas han dado lugar a la necesidad de algoritmos que puedan lidiar con la incorporación de nuevos datos y que no se vean afectados por el orden en el que estos son introducidos.

- **Capacidad de Clusterizar datos con muchos atributos:** Aunque de manera común se suele trabajar con datos que no contienen mucho más que una decena de atributos, hay ciertos tipos de datos como documentos, en los que cada palabra clave puede ser considerada un atributo, dando lugar a cientos de atributos por cada dato, haciendo complicado trabajar con ellos y también clasificarlos, por lo que se precisa de algoritmos que puedan lograr un buen resultado con este tipo de datos.

- **Clustering basado en restricciones:** En el mundo real podemos encontrar ciertas barreras a la hora de aplicar un algoritmo de Clustering, en los que, por ejemplo, sí se tenga que tener en cuenta un límite a la cantidad de cluster que se pueden generar o de decidir si dos datos o clusters deban unirse. Aunque es complicado realizar este tipo de agrupaciones, es indispensable que los algoritmos obtengan resultados decentes dadas estas limitaciones.

- **Interpretabilidad y usabilidad:** De nada sirve clasificar los datos y cumplir con los requisitos previos si no se puede interpretar el resultado. Los algoritmos, por lo tanto, deben ofrecer conclusiones comprensibles, coherentes y que puedan usarse, sometiéndose a ciertas limitaciones semánticas y visuales con la finalidad de acomodar el resultado al objetivo inicial por el que se ha llevado a cabo la clusterización.

### 3.2. Criterios de comparación

[16] Cada algoritmo opera de forma distinta en diferentes aspectos como el criterio de división, la separación de clusters, el cálculo de la similaridad y el espacio de clusterización. Observando los pasos a seguir y la toma de decisiones de un algoritmo, podemos compararlo con otro y escoger cuál es más adecuado para el problema a afrontar.

- **Criterio de división:** Principalmente existen dos tipos de métodos respecto a este criterio, aquellos que fraccionan todos los datos para que no exista ningún tipo de orden o diferencia de nivel entre ellos, y los que, por el lado opuesto, dividen los datos de manera jerárquica, permitiendo formar clusters en diferentes niveles semánticos. Cada uno aporta una utilidad distinta, con el primer tipo de criterio podemos abordar problemas como la clasificación de clientes o lectores, conjuntos de datos



en los todos reciben el mismo trato; mientras que con criterio de partición jerárquico podemos organizar documentos en varios campos como “deportes”, “política” o “comida”, y dentro de cada campo tener subconjuntos como “postres”, “pasta” o “carne”.

- **Separación de clusters:** Otro criterio en el que hay dos enfoques: si un dato forma parte de un cluster, este no puede formar parte de otro a no ser que ambos clusters se unan; o permitir que un dato pueda estar en dos clusters a la vez, un concepto denominado “Fuzzy clustering”. La primera aproximación ofrece un enfoque más determinista que es quizás el más aplicado entre ambos métodos, pero existen disciplinas en las que permitir que un dato forme parte de más de cluster puede ser realmente útil y que cada vez recibe más usos [17].

- **Cálculo de similaridad:** Una de las pautas que más determina un algoritmo de clusterización es la forma en la que calculan la similaridad entre los datos, y que exploraremos detalladamente más adelante. Gran parte de los algoritmos utilizan la distancia que hay entre dos objetos para calcular su similaridad aplicando métodos como la distancia Euclídea o espacios vectoriales. Esta tipo de cálculo se ve beneficiado de métodos de optimización, pero suele generar clusters con formas esférica o cilíndricas, por lo que cada vez más se utilizan algoritmos que miden la similaridad entre objetos mediante la densidad o contigüidad del espacio para formar clusters, lo cual permite descubrir grupos con formas arbitrarias.

- **Espacio de Clusterización:** Muchos algoritmos de Clustering observan todo el espacio de datos para buscar grupos, que es un método totalmente factible y válido para datos con pocos atributos. Sin embargo, cuando se tratan de procesar datos con gran cantidad de atributos, es mejor centrarse en pequeñas partes del total, en subconjuntos de espacio y buscar clusters dentro de ellos de forma que se pueda obtener información valiosa a un menor coste computacional.

### 3.3. Clasificación general de técnicas

[16] Una vez que hemos visto los requisitos de los algoritmos de Clustering y los criterios para compararlos entre sí, podemos agrupar la mayor parte

de las técnicas bajo cuatro grandes grupos basados principalmente en el cálculo de la similaridad y el criterio de división. Hablaremos brevemente sobre estos a continuación y posteriormente, precisaremos de manera individual sobre cada uno de ellos en las secciones 4, 5, 6 y 7, señalando los principales algoritmos dentro de cada una de ellos.

- **Métodos basados en particiones:** Este tipo de métodos dividen todos los objetos de un conjunto de datos en clusters en los que debe haber al menos un objeto; o dicho de una forma más técnica, teniendo  $n$  objetos, se generarán  $k$  particiones siendo  $k \leq n$ .

Normalmente las técnicas de este grupo generan clusters exclusivos, donde un dato solo puede pertenecer a un cluster, pero cada vez se desarrollan más algoritmos de tipo fuzzy que relajan la agrupación. Asimismo, muchos de estos métodos calculan la distancia entre los objetos como medida de similaridad y generan de manera inicial  $k$  particiones para posteriormente, de forma iterativa, tratar de reagrupar los datos y clusters para mejorar la clasificación, basándose en que los datos dentro del mismo cluster estén realmente cerca entre sí, y que con los datos de otros clusters se encuentren lejos.

Estos métodos sirven tanto para búsquedas en todo el espacio, como Clusterización de espacios reducidos, pero cuando la cantidad de atributos es muy alta, es preferible aplicarlos en subconjuntos, pues es complicado optimizar computacionalmente este tipo de algoritmos a la hora de generar las particiones iniciales, y para lidiar con este inconveniente, se suelen usar métodos heurísticos como  $k$ -means que mejoran la calidad de Clustering progresivamente. Sin embargo, como se ha mencionado previamente, esta aproximación de Clustering genera grupos con formas esféricas y circulares, por lo que uno de los principales focos de mejora de estos algoritmos es evitar ese tipo de comportamiento cuando se trabaja con grandes conjuntos de datos.

- **Métodos jerárquicos:** Las técnicas de este grupo crean, con su propio nombre indica, una separación jerárquica de los datos. De forma diferente a los métodos basados en particiones donde todos los datos están en el mismo nivel,

aquí se busca generar varios rangos para la Clusterización.

Existen dos enfoques principales dependiendo de cómo se generen los clusters: *aglomerativo* y *divisivo*. Las técnicas aglomerativas parten de la separación de todos los objetos en clusters individuales y progresivamente los va uniendo hasta que solo queda uno; mientras que las técnicas divisivas parten de un solo cluster inicial en el que se encuentran todos los objetos y va formando clusters más pequeños hasta que cada objeto es en sí un cluster. Es por eso que también se les conoce por métodos de “abajo a arriba” y de “arriba a abajo” respectivamente.

Esta técnicas pueden utilizar distancia, densidad y continuidad como medida de clacular la similaridad entre los datos y son útiles tanto en búsquedas de espacios completos como de subconjuntos.

- **Métodos basados en densidad:** Este tipo de métodos surgieron como solución a los métodos basados en particiones y su tendencia a encontrar únicamente clusters con formas esféricas o circulares, puesto que las técnicas que utilizan la densidad del espacio a la hora de formar clusters permiten descubrir clusters con formas más arbitrarias.

El concepto general de estos métodos consiste en agrandar un cluster mientras la densidad en el “vecindario” este por encima de un umbral. La densidad se calcula por la cantidad de objetos que hay en el espacio de datos y el vecindario hace referencia a todos los datos que se encuentran a cierta distancia de un objeto determinado.

El concepto general es que para cada objeto dentro de un cluster, en su vecindario de radio  $r$  de distancia debe haber un mínimo de  $x$  objetos. Este tipo de aproximación resulta muy conveniente a la hora de detectar datos anómalos.

Los métodos basados en densidad pueden clasificar los datos en clusters exclusivos o de manera jerárquica, y aunque normalmente consideraran que un objeto solo puede pertenecer a un cluster, también existen técnicas fuzzy Clustering. Asimismo, se pueden aplicar tanto en espacios completos como en regiones puntuales.

- **Métodos basados en cuadrícula o rejilla:**

Estos métodos dividen el espacio de objetos en una cantidad determinada de celdas que forman una estructura de cuadrículo o rejilla, de ahí su nombre. Todas las operaciones con los datos se realizan dentro de la rejilla, que permite un procesamiento mucho más rápido, pues no depende de la cantidad de datos, sino de la de celdas existentes.

Esto permite que puedan ser integrados con otros métodos de clustering de otros grupos como jerárquicos y basados en densidad para mejorar su eficiencia.

En la tabla 1 se han resumido estos grupos y las principales características de los mismos. Cabe mencionar que esta clasificación no es perfecta, pues existen algoritmos que integran conceptos de diferentes métodos.

A continuación, exploraremos todos estos grupos y las principales técnicas empleadas en cada uno de ellos.

Table 1: Resumen de métodos de Clusterización

Método	Características generales
Basados en particiones	<ul style="list-style-type: none"><li>- Encuentran clusters de forma esférica mutuamente exclusivos</li><li>- Utilizan distancia/proximidad</li><li>- Usan la media o medoid para representar el centro del cluster</li><li>- Efectivos en conjuntos de datos pequeños y medianos</li></ul>
Jerárquicos	<ul style="list-style-type: none"><li>- Clasifican en múltiples niveles</li><li>- No pueden deshacer agrupaciones o divisiones erróneas</li><li>- Pueden incorporar otras técnicas como microclustering y tener en cuenta vínculos entre los objetos</li></ul>
Basados en densidad	<ul style="list-style-type: none"><li>- Pueden encontrar clusters con formas arbitrarias</li><li>- Los clusters son regiones con gran densidad de objetos separados por zonas con poca densidad</li><li>- La densidad queda definida por un mínimo de objetos cercanos dentro del vecindario</li><li>- Sirve para detectar datos anómalos</li></ul>
Basados en rejilla	<ul style="list-style-type: none"><li>- Utiliza una estructura de rejilla o cuadrícula</li><li>- La velocidad de procesamiento es alta, pues no influye el número de objetos</li></ul>

#### 4. Métodos basados en particiones

Los métodos pertenecientes a este grupo son quizás los más simples y básicos de Clustering. Estos organizan los datos de un conjunto en diversos clusters exclusivos, y es común que soliciten al usuario el número de grupos o clusters finales que debe haber, pues sirve como punto de partida para estos métodos. Esto permite simplificar su funcionamiento, pero a su vez, supone el incumplimiento de uno de los requisitos de los que se ha hablado previamente en la sección 3.1.

De forma técnica, dado un conjunto de datos  $D$  con  $n$  objetos y siendo  $k$  el número de clusters a formar, un algoritmo basado en particiones organizará los objetos en  $k$  divisiones, con  $k \leq n$ , donde cada una de ellas representa un cluster.

Los clusters son generados intentando optimizar el criterio de similaridad, siendo uno de los más utilizados el del cálculo de proximidad, donde se intentan agrupar bajo el mismo clusters aquellos objetos

que estén cerca en el espacio de datos y que a su vez, hacen que todos los datos de un cluster estén alejados de los objetos de otro cluster, asegurando de esta manera un alto grado de semejanza entre los del mismo cluster y diferencia con los de otros clusters.

A continuación se exponen los dos algoritmos más utilizados de este grupo y de los más conocidos:  $k$ -media ( $k$ -means) y  $k$ -medoids.

#### 4.1. $k$ -Means

El objetivo de los algoritmos basados en particiones es, dado un conjunto de datos  $D$  con  $n$  objetos en un espacio Euclídeo<sup>12</sup>, formar  $k$  clusters,  $C_1, \dots, C_k$ , donde cada cluster es un subconjunto del conjunto de datos  $C_i \subset D$  y la intersección entre clusters es nula  $C_i \cap C_j = \emptyset$  para  $i \leq j$  y  $j \leq k$ . Sin embargo, estos requieren de una gran capacidad computacional, y es por ello que algoritmos como  $k$ -Means son tan utilizados.

$k$ -Means representa cada cluster con su “centroide”, que es básicamente el punto central del cluster, y se calcula como la media de los puntos que lo forman. Para un conjunto de datos con dos atributos cuantitativos, la fórmula para calcular el centroide  $c_i$  es la siguiente:

$$c_i = \left( \frac{\sum_{i=1}^n x_i}{n}; \frac{\sum_{j=1}^m y_j}{m} \right). \quad (1)$$

Asimismo, otro de los métodos más comunes para medir la distancia entre puntos y clusters, y que también utiliza  $k$ -Means, es la distancia Euclídea. La fórmula de la distancia Euclídea entre dos puntos cualesquiera  $dist(p, q)$  de un espacio euclídeo  $p$  y  $q$ , con coordenadas cartesianas<sup>13</sup>  $(x_1, y_1)$  y  $(x_2, y_2)$  es como sigue:

$$d(p, q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (2)$$

Y para comprobar la calidad de un cluster, y garantizar que los objetos del un mismo sean similares entre ellos, pero diferentes con objetos de otros clusters, los métodos basados en particiones suelen utilizar la *varianza* dentro del cluster, que es la

suma del error cuadrático<sup>14</sup> entre todos los objetos del cluster  $C_i$  y el centroide  $c_i$ , definida como:

$$E = \sum_{i=1}^k \sum_{p \in C_i} dist(p, c_i)^2, \quad (3)$$

donde  $E$  es la suma del error cuadrático de todos los objetos pertenecientes al cluster  $p_i \in C_i$ . La finalidad de esta operación es asegurar que los  $k$  cluster resultantes sean lo más compactos y distantes posible.

Es aquí, en el cálculo de la varianza donde la optimización de este método se vuelve computacionalmente complicada, pues en el peor de los casos, el total de posibles clasificaciones depende del número total de clusters de manera exponencial y posteriormente habría que comprobar los valores de la varianza dentro de cada cluster, siendo un problema de complejidad NP-complejo<sup>15</sup>. Para resolverlo, se utilizan algoritmos voraces<sup>16</sup> como  $k$ -Means, que es simple y usado habitualmente.

Este algoritmo sigue una serie de pasos para formar los clusters y evaluar su calidad. Primero, selecciona  $n$  objetos dentro del conjunto de datos  $D$  para representar los clusters iniciales; y puesto que son los únicos integrantes de ellos, el cálculo del centroide da como resultado el propio objeto. Posteriormente, asigna cada uno de los objetos restantes al cluster con el que es más similar basándose en el cálculo de la distancia Euclídea entre el objeto y representante del cluster. A partir de aquí, el algoritmo intenta mejorar los resultados de la varianza, para ello, vuelve a calcular el centroide de cada cluster con los nuevos objetos añadidos en la iteración previa, y reagrupa los objetos ahora con los representantes de cada cluster actualizados. El proceso continua hasta que los centroides, y por ende, clusters de la iteración actual sean idénticos a los de la iteración previa.

El pseudocódigo de  $k$ -Means queda resumido en el algoritmo 1.

Sin embargo, este método no garantiza converger hacia un máximo global, sino que muchas veces termina en un máximo local. El resultado depende de

<sup>12</sup>Espacio geométrico donde se satisfacen los axiomas de Euclides.

<sup>13</sup>Tipo de coordenadas ortogonales usadas en espacios Euclídeos que tienen como referencia los ejes ortogonales.

<sup>14</sup>Técnica para calcular la diferencia entre el estimador y lo que se estima.

<sup>15</sup>No se definir esto bien

<sup>16</sup>Estrategia de búsqueda en la que se sigue una heurística que consiste en escoger la opción más optima en cada paso de manera local, esperando llegar a una solución general óptima.



---

**Algoritmo 1:  $k$ -Means**

---

**Entrada:**  $k$ : número de clusters,  
 $D$ : conjunto de datos con  $n$  objetos.  
**Salida:** Un conjunto de  $k$  clusters.

```
1 inicio
2   selección arbitraria de  $k$  objetos de  $D$ 
   como representantes de los clusters
   iniciales
3   repetir
4     calcular la distancia de cada objeto al
       centroide de cada cluster
5     (re)asignar cada objeto al cluster con
       el que es más similar en función de
       la distancia obtenida
6     actualizar los centroides de los
       clusters con los nuevos objetos
       añadidos
7   hasta no hay cambios
8 fin
```

---

pero supone un alto coste computacional.

Además, recordemos que este algoritmo flaquea a la hora de encontrar clusters con formas poco convencionales y de tamaños dispares, siendo realmente sensible a datos anómalos e impurezas, que pueden alterar los cálculos de los centroides.

## 5. Métodos jerárquicos

## 6. Métodos basados en densidad

## 7. Métodos basados en rejilla

la selección de clusters iniciales sobre los que se trabaja, por lo que es común realizar varias pruebas del algoritmo con diferentes clusters iniciales hasta encontrar un resultado óptimo.

Por otro lado, la complejidad temporal de  $k$ -means es considerablemente baja  $O(nkt)$ , pues depende del número de objetos  $n$ , del de clusters  $k$  y de la cantidad de iteraciones  $t$  y normalmente encontramos que  $k \ll n$  y  $t \ll n$ . Lo cual sugiere que escalar este algoritmo a grandes conjuntos de datos es asequible.

Asimismo, este algoritmo solo puede ser utilizado cuando podemos calcular la media de varios objetos para encontrar el centroide, por lo que en casos donde los datos tengan atributos nominales, no se puede aplicar. Pero existen modificaciones y versiones de  $k$ -means que solucionan algunos de estos problemas, como  $k$ -modas, que en vez de utilizar la media para calcular el centroide, utiliza la moda, el atributo que más se repite para una característica, y si permite trabajar con datos nominales.

Otra de las desventajas que presenta y que comentábamos al inicio de la sección, es el hecho de que no cumple uno de los requisitos de los métodos de clustering, y es que  $k$ -means pide al usuario el número  $k$  de clusters a generar como dato de entrada indispensable. Si bien es cierto que se han propuesto estudios para solventar este problema intentando encontrar el mejor valor para  $k$  ejecutando varias veces el algoritmo con distintas unidades,

## 8. Algoritmos

1	<a href="#">k-Means</a>	9
---	-------------------------	---

## 9. Tablas

1	<a href="#">Resumen de métodos de Clusterización</a>	7
---	--	---

## 10. Imágenes

## 11. Referencias

- [1] Alberts, D. S., & Papp, D. S. (1997). [The information age: An anthology on its impact and consequences](#). Office of the Assistant Secretary of Defense Washington DC Command and Control Research Program (CCRP).
- [2] Becoming A Data-Driven CEO — Domo. (2018). Data never sleeps 6.0 <https://www.domo.com/solution/data-never-sleeps-6>
- [3] Xu, Z., & Shi, Y. (2015). [Exploring big data analysis: fundamental scientific problems](#). Annals of Data Science, 2(4), 363-372.
- [4] Definición Data Science apuntes FCD
- [5] Definición Data Mining libro 100
- [6] The R Project for Statistical Computing. (n.d.). Retrieved from <https://www.r-project.org/>
- [7] Moreno, A. (1994). [Aprendizaje automático](#). Llibre, Edicions UPC.
- [8] Apuntes JJ clustering
- [9] Alkhaibari, A. A., & Chung, P. (2017). [Cluster analysis for reducing city crime rates](#). 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT). doi:10.1109/lisat.2017.8001983
- [10] Song, Y., Meng, H., & Zhang, Y. (2010). [Clustering analysis and its applications](#). 2010 Second IITA International Conference on Geoscience and Remote Sensing. doi:10.1109/iita-grs.2010.5602787
- [11] Prabhu, J., Sudharshan, M., Saravanan, M., & Prasad, G. (2010). [Augmenting Rapid Clustering Method for Social Network Analysis](#). 2010 International Conference on Advances in Social Networks Analysis and Mining. doi:10.1109/asonam.2010.55
- [12] Baron, J. N., Aznar, M. N., Monterubbianesi, M., & Martínez-López, B. (2020). [Application of network analysis and cluster analysis for better prevention and control of swine diseases in Argentina](#). PLoS ONE, 15(6), 1–26. <https://doi.org/10.1371/journal.pone.0234489>

- [13] Li, J., & Chen, P. (2008). [The application of Cluster analysis in Library system](#). 2008 IEEE International Symposium on Knowledge Acquisition and Modeling Workshop. doi:10.1109/kamw.2008.4810639
- [14] Emebo, O., Aromolaran, O., Oyelade, J., Isewon, I., Oladipupo, O., Omogbadegun, Z., ... Olawole, O. (2019). Data Clustering: Algorithms and Its Applications. 2019 19th International Conference on Computational Science and Its Applications (ICCSA). doi:10.1109/iccsa.2019.000-1
- [15] Carugo, O., & Eisenhaber, F. (2010). Data Mining Techniques for the Life Sciences. Humana Press.
- [16] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed., p. 740). 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann Publishers, Elsevier.
- [17] Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). Fuzzy cluster analysis: methods for classification, data analysis and image recognition. John Wiley & Sons.