



TRABAJO FIN DE GRADO  
INGENIERÍA INFORMÁTICA

# Clustering Con Restricciones

---

Un Enfoque Práctico

**Autor**

Germán González Almagro

**Directores**

Julián Jesús Luengo Navas  
Salvador García López



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE  
TELECOMUNICACIÓN

—  
Granada, mes de 201



# Clustering Con Restricciones

---

Un Enfoque Práctico.

**Autor**

Germán González Almagro

**Directores**

Julián Jesús Luengo Navas  
Salvador García López

# **Clustering Con Restricciones: Un Enfoque Práctico**

Germán González Almagro

**Palabras clave:** palabra\_clave1, palabra\_clave2, palabra\_clave3, .....

## **Resumen**

Poner aquí el resumen.

---

Yo, **Germán González Almagro**, alumno de la titulación en ingeniería informática de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 76593910T, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Germán González Almagro

Granada a X de mes de 201 .



---

D. **Julián Jesús Luengo Navas**, Profesor del Área de XXXX del Departamento Ciencias de la Computación en Inteligencia Artificial de la Universidad de Granada.

D. **Salvador García López**, Profesor del Área de XXXX del Departamento Ciencias de la Computación en Inteligencia Artificial de la Universidad de Granada.

**Informan:**

Que el presente trabajo, titulado *Clustering Con Restricciones, Un Enfoque Práctico*, ha sido realizado bajo su supervisión por **Germán González Almagro**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a X de mes de 201 .

**Los directores:**

Julián Jesús Luengo Navas      Nombre Salvador García López



## AGRADECIMIENTOS

---

Poner aquí agradecimientos...



## ÍNDICE GENERAL

---

<b>1</b>	<b>INTRODUCCIÓN</b>	<b>1</b>
1.1	Motivación Personal . . . . .	2
1.2	Objetivos . . . . .	2
<b>2</b>	<b>BREVE INTRODUCCIÓN AL CLUSTERING</b>	<b>4</b>
2.1	Motivación para la clasificación . . . . .	4
2.2	Métodos numéricos para el Clustering . . . . .	4
2.2.1	¿Qué es un cluster? . . . . .	6
2.3	Aplicaciones del clustering . . . . .	8
2.3.1	Aplicaciones en marketing . . . . .	8
2.3.2	Aplicaciones en astronomía . . . . .	9
2.3.3	Aplicaciones en psiquiatría . . . . .	9
2.3.4	Aplicaciones en meteorología y climatología . .	10
2.3.5	Aplicaciones en arqueología . . . . .	10
2.3.6	Aplicaciones en bioinformática y genética . . .	11
2.4	Resumen . . . . .	12
<b>3</b>	<b>CLUSTERING CON RESTRICCIONES</b>	<b>13</b>
3.1	Definición de las restricciones . . . . .	13
3.2	Uso de las restricciones . . . . .	15
3.2.1	Métodos basados en restricciones . . . . .	15
3.2.2	Métodos basados en distancia . . . . .	16
3.3	Aplicaciones del clustering con restricciones . . .	17
3.3.1	Aplicaciones en análisis de imágenes . . . . .	17
3.3.2	Aplicaciones en análisis de vídeos . . . . .	19
3.3.3	Aplicaciones en genética . . . . .	19
3.3.4	Aplicaciones en análisis de textos . . . . .	20
	<b>BIBLIOGRAFÍA</b>	<b>21</b>

## ÍNDICE DE FIGURAS

---

Figura 2.1	Clusters con cohesión interna y/o aislamiento externo . . . . .	7
Figura 2.2	Distribución uniforme de puntos . . . . .	7
Figura 2.3	Distribución uniforme de puntos clasificados . . . . .	7
Figura 3.1	Restricciones de tipo delta y epsilon. [1] . . . . .	14
Figura 3.2	Restricciones sobre un conjunto de datos [1] . . . . .	16
Figura 3.3	Clustering que satisface todas las restricciones [1] . . . . .	16
Figura 3.4	Restricciones sobre un conjunto de datos [1] . . . . .	17
Figura 3.5	Clustering basado en métrica aprendida en base a las restricciones [1] . . . . .	17
Figura 3.6	Caras de la base de datos de Carnegie Mellon University (CMU) [1] . . . . .	17
Figura 3.7	Restricciones de tipo Cannot-Link (CL) entre caras de la misma persona [1] . . . . .	18
Figura 3.8	Método de clustering empleado en el sistema de navegación del robot Aibo [1] . . . . .	18
Figura 3.9	Diferentes tipos de restricciones en datos de video [1] . . . . .	19
Figura 3.10	Clustering de genes basado en microarrays [1]	20

## ÍNDICE DE CUADROS

---

## LISTINGS

---

## ACRONYMS

---

ML      Must-Link

CL      Cannot-Link

CMU   Carnegie Mellon University



## INTRODUCCIÓN

---

*An intelligent being cannot treat every object it sees as unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand.*

**Steven Pinker, How the Mind Works, 1997**

Una de las habilidades más básicas y primitivas de la que están dotados los seres humanos es la de agrupar objetos similares para producir una clasificación que les resulte útil. Habilidad que ya nuestros más antiguos ancestros debieron poseer, por ejemplo, debieron ser capaces de darse cuenta de qué objetos eran comestibles, cuales eran venenosos y cuales intentarían matarles.

La capacidad de clasificación, en su sentido mas amplio, es necesaria para el desarrollo del lenguaje, que esta formado por palabras que nos ayudan a reconocer diferentes tipos de eventos, acciones y entidades. En esencia, cada sustantivo es una etiqueta que empleamos para agrupar un colectivo de seres u objetos con características similares, de manera que podemos hacer referencia a todos ellos empleando la palabra que los une.

De igual forma que la clasificación es una habilidad básica para las personas en su vida cotidiana, es también esencial en la mayoría de las ramas de la ciencia. En biología, por ejemplo, la clasificación de los diferentes tipos de organismos ha sido objeto de estudio desde el comienzo de su existencia. Aristóteles construyó un elaborado sistema de clasificación animal que dividía todas a las criaturas en dos grupos, aquellos con sangre roja y aquellos que carecían de ella. Más tarde propuso una subdivisión que los clasificaba según la forma en la que nuevos individuos venían al mundo, ya sea vivos, mediante huevos, crisálidas, etc.

Siguiendo a Aristóteles, Teoprastos escribió el primer documento que recopilaba las directrices para la clasificación de las plantas. Los trabajos resultantes fueron tan amplios y detallados que sentaron las bases para la investigación en biología durante los siguientes siglos. Este trabajo no fue sustituido hasta 1737, cuando Carlos Linneo publicó su trabajo *Genera Plantarum*, del que el siguiente fragmento ha sido extraído:

*All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar. The greater the number of natural distinctions this method comprehends the clearer becomes our idea of things. The more nume-*

*rous the objects which employ our attention the more difficult it becomes to form such a method and the more necessary. For we must not join in the same genus the horse and the swine, though both species had been one hoof'd nor separate in different genera the goat, the reindeer and the elk, tho' they differ in the form of their horns. We ought therefore by attentive and diligent observation to determine the limits of the genera, since they cannot be determined a priori. This is the great work, the important labour, for should the genera be confused, all would be confusion.*

**Carlos Linneo, Genera Plantarum, 1737**

La clasificación de los animales y las plantas ha jugado un importante papel en campos como la biología y la zoología. Particularmente, esta clasificación sentó las bases para el desarrollo de la teoría de la evolución de Darwin. Pero también ha sido de gran relevancia en áreas de conocimiento como la química y la física, con la clasificación de los elementos en la tabla periódica, propuesta por Mendeleyev en la década de 1860; o en astronomía, con la clasificación de estrellas en enanas o gigantes empleando las directrices de Hertzsprung–Russell.

#### MOTIVACIÓN PERSONAL

Durante mi formación he escuchado de multitud de profesores que incorporar conocimiento humano a una máquina es una de las tareas mas complejas a las que se ha enfrentado la humanidad.

Algo que realmente me resultó interesante fue que Deep Blue, la primera máquina en ganar al campeón del mundo de ajedrez, Gary Kasparov en aquel momento (1996), no ganó por un avance significativo en el algoritmo que ejecutaba la máquina, sino por avances en el hardware que permitían que la máquina analizase mas movimientos por unidad de tiempo, es decir, la máquina no empleaba conocimiento del que no disponía anteriormente, simplemente “pensaba” más rápido.

Por ello, he querido profundizar más en el campo de la incorporación de conocimiento a las máquinas algo más de lo que he podido hacerlo durante estos años. El clustering con restricciones puede verse como un ejemplo de ello, al fin y al cabo no es más que guiar el proceso de toma de decisiones de una máquina incorporando conocimiento extraído de las personas.

#### OBJETIVOS

Este documento pretende presentar el clustering con restricciones desde un punto de vista práctico, exponiendo los casos en los que su empleo mejora con creces a las técnicas anteriores, así como los casos en los que no resulta de verdadera utilidad.

En la segunda sección de este documento se da una breve introducción a los conceptos y aplicaciones del clustering sin restricciones, mientras que la sección tres profundizará en el clustering con restricciones, exponiendo algunas de las técnicas que han surgido para su aplicación. Por último, la sección 4 expondrá los resultados obtenidos al experimentar con las técnicas expuestas en la 3.

# 2

## BREVE INTRODUCCIÓN AL CLUSTERING

---

En primer lugar, será necesario realizar una introducción, siquiera sea breve, al clustering y sus aplicaciones, que facilite la comprensión de los siguientes (apartados). A tal fin se seguirá el estudio realizado por Brian S. Everitt et al. (2011) [2].

### MOTIVACIÓN PARA LA CLASIFICACIÓN

La clasificación puede ser entendida como una forma de simplificar la información contenida en grandes conjuntos de datos, de un modo que sea fácilmente comprensible por las personas. De esta manera, los procesos de extracción de información útil y aplicación de la misma se simplifican. Así, si somos capaces de dividir de forma válida un gran conjunto de datos en subconjuntos o grupos, podremos extraer información común a todos los elementos del subconjunto y proporcionar una descripción precisa de los englobes.

La necesidad de analizar la información de esta manera crece con la aparición y disponibilidad de grandes conjuntos de datos en el ámbito de la ciencia. El análisis de este tipo de información mediante clasificación, o clustering, hoy en día es conocido como *Ciencia de datos*. En el siglo 21 surge un particular interés en la ciencia de datos desde la aparición de la *World Wide Web*, conocida como Internet, donde el objetivo se ha convertido en extraer información relevante de las páginas Web que forma esta basta red.

Es importante destacar que en la mayoría de las ocasiones no hay un sólo criterio de clasificación para un mismo conjunto de datos, de hecho existe una amplia variedad de los mismos. En el caso de las personas, podrían ser clasificadas, por ejemplo, en base a sus ingresos económicos, o según la cantidad de calorías que consumen a lo largo de un periodo de tiempo definido. Así, distintos criterios de clasificación no tiene porque dar como resultado la misma división en grupos del conjunto a clasificar, de esta manera, diferentes criterios servirán a diferentes propósitos.

### MÉTODOS NUMÉRICOS PARA EL CLUSTERING

Los métodos numéricos para el clustering surgen en ramas de las ciencias naturales, como la biología o la zoología, en un intento de eliminar la subjetividad implícita en el proceso de clasificación que se desencadena al descubrir una nueva especie. El objetivo es pro-

porcionar un método no subjetivo y estable para clasificar y agrupar elementos.

Estos métodos adoptan diversos nombres que varían según el campo en el que se apliquen, taxonomía numérica (*numerical taxonomy*) en biología, *Q* análisis (*Q analysis*) en psicología, o reconocimiento de patrones no supervisado (*unsupervised pattern recognition*) en el campo de la inteligencia artificial. No obstante, hoy en día, análisis de clusters (*Clusters analysis*) o simplemente clustering son los términos más ampliamente aceptados y extendidos para referirse a tareas que involucran el descubrimiento de subgrupos dentro de un conjunto de elementos.

En la mayoría de las aplicaciones del clustering, el objetivo es obtener una partición de los datos, en la que cada instancia u objeto pertenezca a un único cluster, y la unión de todos ellos contenga a todos los objetos individuales. Dicho esto, debe destacarse que en algunas circunstancias son aceptables soluciones en las que existe solapamiento entre clusters, así como el hecho de que puede no existir una partición aceptable de los datos.

La manera más ampliamente extendida de representar la información sobre la que se debe aplicar clustering es una matriz  $X$  de dimensión  $n \times p$ , en la que cada fila corresponde a una instancia u objeto a procesar, y cada columna corresponde a una de las variables que caracterizan dichas instancias u objetos. El término comúnmente aceptado para referirse a cada fila es el de *vector de características*.

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & \cdots & x_{n,p} \end{pmatrix}$$

La entrada  $x_{i,j}$  en  $X$  se corresponde con el valor de la variable  $j$ -esima en la instancia  $i$ .

Las variables en  $X$  pueden ser una mezcla de atributos en un dominio continuo, discreto o categórico. Además, es posible que, en problemas reales, algunas entradas no estén disponibles. Esta mezcla de tipos de variables y los valores perdidos pueden complicar la tarea del clustering, sin embargo, existen métodos para tratar estos casos, como la inferencia de valores perdidos o las transformaciones de dominio.

En algunas aplicaciones, las filas de la matriz pueden contener medidas repetidas de la misma variable, aunque bajo diferentes condiciones, o en diferentes momentos, incluso en diferentes localizaciones espaciales. Un ejemplo de ello pueden ser las medidas de altura de un grupo de niños en un mismo mes a lo largo de diferentes años. Este tipo de datos poseen una estructura que, de nuevo, puede complicar la tarea del clustering.

Algunos métodos de clustering conllevan realizar transformaciones sobre la matriz  $X$  para transformarla en una matriz de  $n \times n$  en la que se almacenan medidas extraídas de la matriz  $X$  que relacionen una instancia con todas las demás, como pueden ser la similitud, distancia o disimilitud.

El clusterign es, dicho de manera simple, descubrimiento de grupos en datos, y no debe ser en ningún caso confundido con los métodos de discriminación o asignación, conocidos en el ámbito de la inteligencia artificial como aprendizaje supervisado, en los que los grupos son conocidos a priori y el objetivo del análisis es obtener una regla de clasificación o clasificador que permita asignar nuevas instancias o individuos a uno de los grupos ya conocidos.

Una vez definida la estructura general de los métodos de clustering, estaría justificado preguntar, ¿qué es un cluster? El siguiente epígrafe intentará dar respuesta a esta pregunta.

### *¿Qué es un cluster?*

Hasta este momento, los términos cluster, grupo y clase han sido empleados de una manera completamente intuitiva, sin necesidad alguna de definición formal, una prueba más de lo innato de estos conceptos en el ser humano. De hecho, dar una definición formal de cluster resulta una tarea, no solo complicada, sino en muchas ocasiones poco útil. Bonner, por ejemplo, en 1964 propuso una definición de cluster completamente dependiente de la interpretación del usuario, en lo que a él respecta, un cluster es aquello que el usuario entiende como cluster sin haberle propuesto una definición formal del mismo.

Aunque la definición de Bonner es acertada en una amplio rango de situaciones, autores como Cormack, en 1971, o Gordon en 1999, proponen una definición algo más analítica desde el punto de vista matemático, definiendo un cluster en términos de cohesión interna (homogeneidad), y aislamiento externo (separación). La figura 2.1 muestra, de manera informal, las propiedades descritas anteriormente, de forma que, a cualquier observador, le resultarán aparentes los clusters presentes en ella, sin necesidad de una definición formal de cluster. Este hecho puede explicar porqué alcanzar una definición matemáticamente precisa de homogeneidad y separación puede llegar a ser innecesario.

No queda completamente clara la manera en que las personas reconocen diferentes clusters cuando estos son representados en un plano, pero una de las variables que con certeza influye es la distribución de distancias relativas entre los objetos o puntos.



Figura 2.1: Clusters con cohesión interna y/o aislamiento externo

Por otra parte, como ya mencionamos anteriormente en esta sección, pueden existir conjuntos de datos en los que no exista una partición justificada. En la figura 2.2 se muestra un conjunto de datos para el que la mayoría de observadores llegaría a la conclusión de que no existen grupos diferenciados, simplemente una nube de puntos uniformemente distribuida. Idealmente, es de esperar que un método de clustering aplicado a este mismo conjunto de datos llegue a la misma conclusión.

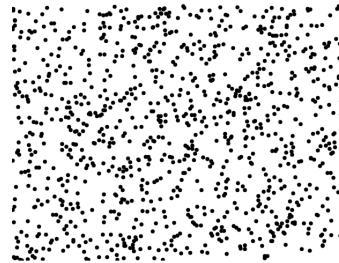


Figura 2.2: Distribución uniforme de puntos

Sin embargo, la mayoría de métodos de aprendizaje no supervisado darán como resultado un particionamiento uniforme como el que se muestra en la figura 2.3. El número de particiones encontrado dependerá del método aplicado, si bien en cualquier caso obtendremos un particionamiento uniforme.

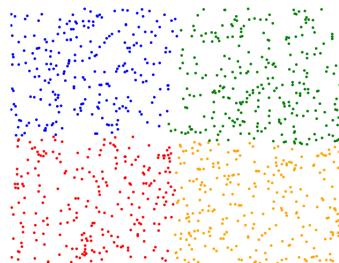


Figura 2.3: Distribución uniforme de puntos clasificados

El proceso de dividir una distribución homogénea de datos en diferentes grupos se conoce como disección, y tal proceso puede ser útil en ciertas circunstancias. Sin embargo, dado que en la mayoría de las ocasiones de aplicación real de métodos de clusters no se conoce a priori la estructura de los datos, existe el riesgo de interpretar todas

las soluciones en términos de existencia de subgrupos, lo que conllevaría a la imposición de una estructura ficticia en datos en los que no hay estructura presente.

#### APLICACIONES DEL CLUSTERING

Como ya se ha indicado, el problema general al que el intenta dar solución el clustering está presente en muchas disciplinas de la ciencia: biología, botánica, medicina, psicología, geografía, marketing, procesamiento de imágenes, psiquiatría, arqueología, etc. En esta sección se presentan algunas de las aplicaciones del clustering relacionadas con las citadas disciplinas.

##### *Aplicaciones en marketing*

Dividir los clientes en grupos homogéneos es una de las tareas más frecuentes en marketing. Un director de marketing podría preguntarse como agrupar los posibles clientes según los beneficios potenciales del producto que intenta introducir en el mercado. Por otra parte, un analista de marketing podría estar interesado en agrupar las empresas según sus características financieras, para poder analizarlas y predecir sus estrategias de mercado.

Un ejemplo de aplicación del clustering en este ámbito fue publicado en Green et al. (1967). Así, con un gran número de ciudades disponibles para el análisis, debieron restringir los lugares en los que llevar a cabo sus estudios, debido a motivos económicos. Para ello hicieron uso del análisis de clusters, clasificando las ciudades en pequeños grupos basándose en 14 características de las mismas, entre ellas se encontraban el tamaño e ingresos medios *per capita*. Dado que se esperaba que las ciudades incluidas en un mismo grupo fueran muy similares, escogieron una ciudad de cada uno de ellos para realizar sus estudios.

Otra aplicación del análisis del clusters en el marketing fue descrita por Chakrapani (2004). En este caso, un fabricante de coches cree que comprar un coche deportivo no es una decisión basada sólo en capacidades económicas o edad, sino que es una decisión relacionada con el estilo de vida que llevan aquellos que deciden comprar un coche de estas características, frente a aquellos que no lo hacen. Consecuentemente, el fabricante decide realizar un estudio, empleando análisis de clusters, que le permita identificar todas las características relacionadas con las personas que comprarían un coche deportivo, para así enfocar sus campañas de marketing a este sector específicamente.

### *Aplicaciones en astronomía*

Dado un conjunto de datos astronómicos, los investigadores quieren saber, por ejemplo, cuantas clases de estrellas hay presentes en ellos, basándose en algún criterio estadístico. Las preguntas más frecuentes dentro de este ámbito son: ¿Cuantos objetos estadísticamente diferentes están presentes en los datos y a que clase debe ser asignado cada objeto? ¿Aparecen clases de objetos previamente desconocidas?. El análisis de clusters puede ser aplicado para dar respuesta a estas cuestiones, ayudando a detectar objetos estadísticamente anómalos, así como a guiar el proceso de clasificación de los mismos. Algunos ejemplos incluyen el descubrimiento de quasars con alto corrimiento al rojo, quasars de tipo 2 (altamente luminosos, núcleos galácticos activos a menudo oscurecidos por polvo y gas), y enanas marrones.

Un ejemplo específico viene dado por el estudio de Faúndez-Abans et al. (1996), que aplicó técnicas de clustering a datos sobre la composición química de 192 nebulosas planetarias. Se identificaron 6 grupos diferentes que eran similares en muchos aspectos a una clasificación previa de dichos objetos, pero que también mostraban diferencias interesantes que hasta ese momento los investigadores había pasado por alto.

Un segundo ejemplo lo encontramos en el estudio de Celeux y Govaert (1992), quienes aplicaron cluster basado en distribuciones normales a un conjunto de 2370 estrellas, descritas por su velocidad relativa al núcleo galáctico y a la rotación galáctica. Usando un modelo de tres clusters, encontraron un cluster de gran tamaño y pequeño volumen, y dos de pequeño tamaño y gran volumen.

### *Aplicaciones en psiquiatría*

Las enfermedades de la mente son a menudo más difíciles de diagnosticar que las enfermedades del cuerpo, es por ello que en el campo de la psiquiatría ha crecido el interés por las técnicas de análisis de clusters que permitan refinar, o incluso redefinir, las técnicas de diagnosis para este tipo de enfermedades. Gran parte de este trabajo involucra pacientes deprimidos, en los que el interés reside en distinguir entre dos tipos de depresión, la endógena (congénita), y la neurótica.

Pilowsky et al. (1968), por ejemplo, usando métodos desarrollados por otros autores, aplicó técnicas de clustering a 200 pacientes en base a sus respuestas a un cuestionario sobre la depresión, junto a información sobre su sexo, edad, estado mental y enfermedad padecida. Este es un claro ejemplo de variables de diferentes tipos incluidas en el mismo conjunto de datos. Uno de los grupos obtenidos como resultado de este estudio fue identificado como marcador de la depresión endógena.

El análisis de clusters también ha sido empleado para encontrar una clasificación de individuos que intentaron cometer suicidio, que podría sentar las bases para estudios posteriores sobre las causas y tratamientos del problema. Paykey y Rassaby (1978), por ejemplo, estudiaron 236 casos de suicidas fallidos registrados por el servicio de emergencias de una ciudad de los Estados Unidos de América. Del conjunto de las variables posibles, 14 fueron seleccionadas como particularmente relevantes para la clasificación, y por tanto fueron usadas en el análisis. Entre ellas se encontraban la edad, número de intentos de suicidio, gravedad de la depresión, grado de hostilidad, además de una serie de características demográficas. Al conjunto de datos resultante se le aplicaron métodos de clustering, el resultado más significativo obtenido corresponde a una división en tres clusters bien definidos.

#### *Aplicaciones en meteorología y climatología*

Diariamente se recogen enormes cantidades de datos sobre la meteorología mundial, explorar estos datos mediante técnicas de clustering puede aportar nuevos enfoques para la climatología y el medio ambiente.

Littmann (2000), por ejemplo, aplicó clustering a los datos recogidos sobre los cambios diarios en la presión superficial en la cuenca Mediterránea, y encontró 20 grupos que explicaban la varianza de las lluvias en las regiones centrales del Mediterráneo. Otro ejemplo viene de la mano de Liu y George (2005), quienes usaron el algoritmo *fuzzy k-means* a datos espaciotemporales de la climatología de las regiones del sur central de EEUU.

#### *Aplicaciones en arqueología*

La arqueología es otra de las disciplinas en la que resulta útil la aplicación del clustering. La clasificación de los diferentes objetos encontrados en los yacimientos puede ayudar a descubrir su uso, los períodos a los que pertenecen, así como la población que los utilizó. De forma similar, el estudio de materiales fosilizados puede ayudar a revelar cómo vivieron las sociedades prehistóricas.

Un ejemplo temprano de la aplicación de clustering a objetos arqueológicos viene dado por Hodson et al. (1966), que aplicó técnicas de clustering a un grupo de broches que datan de la Edad de Hierro, encontrando una clasificación para los mismos de demostrada relevancia arqueológica. Otro ejemplo de la mano de Hodson (1971) es la aplicación del algoritmo *k-means* para construir una taxonomía de hachas de mano encontradas en las Islas Británicas. Las variables tenidas en cuenta para describir cada hacha incluyen longitud, anchura y valores en una escala que describen cómo de puntiaguda era la

herramienta. El clustering dio como resultado dos grupos de hachas, uno formado por las pequeñas y delgadas, y otro formado por las grandes y gruesas.

Respecto a materiales fosilizados, Sutton y Reinhard (1995) realizaron un estudio sobre 155 coprolitos encontrados en *Antelope House*, un yacimiento prehistórico en el Cañón de Chelly en Arizona. El estudio arrojó como resultado una interpretación de las diferencias entre coprolitos basada en la alimentación.

#### *Aplicaciones en bioinformática y genética*

Tiempos recientes están siendo testigo de un tremendo crecimiento en el interés por la Bioinformática, acompañada por la biología molecular, ciencias de la computación, matemáticas y estadística. Tal crecimiento ha sido acelerado por la siempre creciente base de datos genómica y proteica, que son por sí mismas resultado de un grandísimo avance en las técnicas de secuenciación del ADN, medidas de expresión de los genes y compresión de las estructuras macromoleculares. La estadística ha jugado un papel relevante en el estudio de la expresión de los genes. Los genes contenidos en el ADN de cada célula proporcionan las plantillas necesarias para la generación de las proteínas implicadas en la mayoría de los procesos estructurales y biomecánicos que tienen lugar en cada uno de nosotros. Sin embargo, aunque la mayoría de las células en los seres humanos contiene todos los complementos genéticos que componen el genoma humano, los genes se expresan de manera selectiva en cada célula dependiendo del tipo de la misma, del tejido y de las condiciones generales tanto dentro como fuera de la célula. La biología molecular ha puesto de manifiesto que la mayoría de los procesos en la vida de una célula están regulados por factores que afectan a la expresión de sus genes.

Como hemos visto, uno de los campos de investigación más activos hoy en día es el que estudia los procesos que regulan la expresión de los genes. Con el fin de almacenar la información relativa a esta área de estudio surgen los microarrays, (Cortesse, 2000). Desde el punto de vista del análisis de datos, una de las características relevantes en este tipo de información es que el número de características de cada instancia ( $p$ ), supera con creces al número de instancias disponibles ( $n$ ); conjuntos de datos como este son calificados como *datos de alta dimensionalidad*.

La mayoría de métodos estadísticos clásicos no pueden ser aplicados a este tipo de conjuntos de datos sin ser modificados de forma sustancial. Sin embargo, el análisis de clusters acepta bien tales conjuntos de datos y puede ser empleado para identificar grupos de genes con patrones de expresión similares, y dar respuesta a preguntas como porqué un gen se ve afectado por cierta enfermedad, o qué genes son responsables de enfermedades genéticas hereditarias.

Un ejemplo de aplicación lo encontramos en el trabajo de Selinski e Ickstadt (2008), quienes usaron clustering sobre polimorfismos de nucleótidos simples para detectar diferencias entre enfermedades a nivel genético.

#### RESUMEN

Las técnicas de clustering consisten en la exploración de conjuntos de datos sobre los que se debe discernir si pueden o no ser resumidos de manera significativa en términos de un número relativamente pequeño de grupos o clusters de objetos o individuos que se parecen unos a otros y que se diferencian de los que se encuentran en otros clusters.

Muchas ramas de la ciencia han hecho uso de las técnicas de clustering, de manera exitosa, para avanzar en sus respectivos campos y procesar grandes cantidades de datos, cuyo análisis sería impensable afrontar con otras técnicas.

# 3

## CLUSTERING CON RESTRICCIONES

---

Tal y como hemos estudiado en epígrafes anteriores, los métodos de clustering no supervisado son útiles para dotar de estructura a datos referentes a un área concreta. Un ejemplo de ello lo encontramos en la clasificación de textos; Cohn et al. (2003) afrontan un problema propuesto por Yahoo!, que consiste en, dada una gran cantidad de documentos de texto, agruparlos según una taxonomía en la que los documentos con temáticas similares se encuentren cercanos. Para ello, los métodos de clustering no supervisado resultan de utilidad, ya que la información sobre el problema de la que se dispone inicialmente es limitada. Sin embargo, Wagstaff et al. (2001) mostraron que aplicando clustering no supervisado a ciertos problemas, como el de agrupar datos de GPS de forma que los clusters definan los carriles de una vía, no se obtienen resultados significativos, pues los clusters obtenidos distan mucho de la forma alargada que se esperaría como resultado. Para atajar el problema, introdujeron en el clustering un nuevo elemento, las restricciones a nivel de instancia, que permitían incluir conocimiento sobre los clusters que guiarían los métodos de clustering para obtener los resultados esperados. Bastaba con indicar que los carriles de la vía por la que circulan los vehículos miden cuatro metros de ancho, y por tanto cualquier vehículo que se encuentre a una distancia mayor de 4 metros de otro, en dirección perpendicular a la del desplazamiento, debe ser ubicado en un cluster diferente.

Nos situamos entonces en un nuevo escenario: es posible incorporar información adicional al proceso de clustering, además de la contenida en el propio conjunto de datos, para guiarlo en la formación de la partición y obtener resultados más precisos. Esto sitúa al clustering con restricciones en el marco del aprendizaje semisupervisado, a diferencia de los métodos de clustering tradicionales que se enmarcan en el área del clustering no supervisado.

### DEFINICIÓN DE LAS RESTRICCIONES

El nuevo tipo de información que incorporamos al clustering viene dada en forma de restricciones a nivel de instancia, esto es, especificar si dos instancias del conjunto de datos deben estar en el mismo cluster, o, por el contrario, deben estar en clusters separados.

A las restricciones que indican que dos puntos deben ser situados en el mismo cluster se las denomina Must-link, y se notan por  $ML(x, y)$ , donde  $x$  e  $y$  son dos instancias del conjunto de datos. De

manera similar, a las restricciones que especifican lo contrario se las denomina Cannot-link, y se notan por  $CL(x, y)$ .

Aunque pueden parecer simples, las restricciones definidas de la anterior forma poseen propiedades interesantes. Las Restricciones de tipo Must-link son un ejemplo de relación de equivalencia, y por tanto son simétricas, reflexivas y transitivas, formalizando: Hablando sobre

**Observación 3.1** *Las restricciones de tipo ML son transitivas.* Sean  $CC_i$  y  $CC_j$  componentes conexas, conectadas mediante restricciones ML, y sean  $x$  e  $y$  dos instancias en  $CC_i$  y  $CC_j$  respectivamente. Entonces  $ML(x, y) : x \in CC_i, y \in CC_j \rightarrow ML(a, b) \forall a, b : a \in CC_i, b \in CC_j$

**Observación 3.2** *Las restricciones de tipo CL pueden ser (encadenadas).* Sean  $CC_i$  y  $CC_j$  componentes conexas, conectadas mediante restricciones ML, y sean  $x$  e  $y$  dos instancias en  $CC_i$  y  $CC_j$  respectivamente. Entonces  $CL(x, y) : x \in CC_i, y \in CC_j \rightarrow CL(a, b) \forall a, b : a \in CC_i, b \in CC_j$

Un claro ejemplo de uso de las restricciones lo encontramos en casos de aplicación de clustering en los que existen restricciones en cuanto a las medidas de distancia, como sucedía en el supuesto de los datos GPS. Por ejemplo, si queremos que las instancias que forman dos clusters estén separadas por una distancia mayor o igual a  $\delta$ , basta con establecer restricciones de tipo ML entre todas aquellas instancias cuya distancia sea menor que  $\delta$ . De manera similar, si queremos que el diámetro de los clusters sea como mucho  $\epsilon$ , debemos establecer un conjunto de restricciones de tipo CL entre todas aquellas instancias que se encuentren a una distancia mayor que  $\epsilon$ . La figura 3.1 muestra una representación gráfica de estos dos tipos de restricciones.

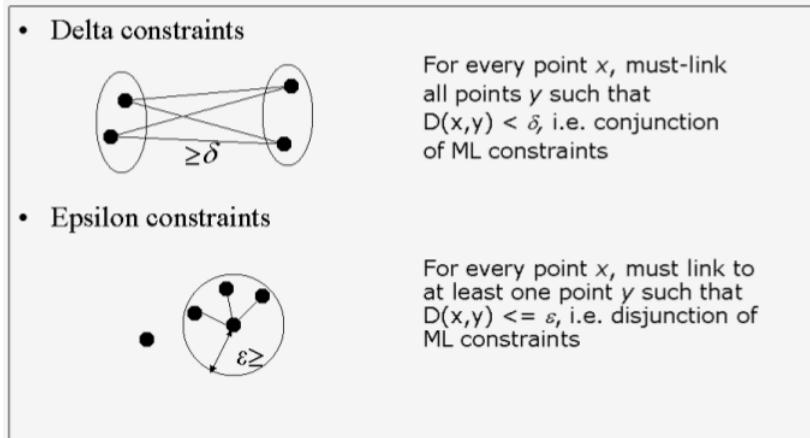


Figura 3.1: Restricciones de tipo delta y epsilon. [1]

## USO DE LAS RESTRICCIONES

Mientras que el aprendizaje completamente supervisado implica conocer la etiqueta asociada a cada instancia, en el aprendizaje semi-supervisado solo se dispone de un subconjunto de instancias etiquetadas. Por otra parte, en gran cantidad de dominios la información disponible se refiere a relaciones entre instancias, y no a la clase concreta a la que pertenecen las mismas. Es más, en montajes de clustering interactivo, un usuario no experto en el dominio del problema podrá, probablemente, aportar información en forma de restricciones de tipo Must-Link (ML) y Cannot-Link (CL), antes que aportar información sobre a qué clase concreta pertenecen ciertas instancias.

Habitualmente, las restricciones se incluyen en los problemas de clustering de dos maneras. Pueden ser empleadas para modificar la regla de asignación de instancias a cluster del método en cuestión, de forma que la solución satisfaga el máximo número de restricciones posible. Alternativamente, cabe la posibilidad de entrenar la función de distancia empleada por el método en base a las restricciones, ya sea antes o durante la aplicación del mismo. En cualquier caso, la fase inicialización puede tomar en consideración las restricciones, de forma que las instancias asociadas con restricciones Must-Link (ML) serán situadas en el mismo cluster, y aquellas entre las que exista una restricción Cannot-Link (CL), quedarán en clusters diferentes. Basándonos en esta distinción, identificamos dos maneras de aproximar el problema, las basadas en restricciones (*constraint-based*), y las basadas en distancias (*distance-based*).

### *Métodos basados en restricciones*

En los métodos basados en restricciones, el propio método de clustering es modificado de manera que la información disponible se emplea para sesgar la búsqueda y obtener una partición de los datos apropiada.

Existen dos modelos de métodos basados en restricciones: (1) aquellos que fuerzan el cumplimiento de las restricciones, e intentan encontrar la mejor asignación posible que no inflaja ninguna de ellas, y (2) las que hacen una interpretación relajada de las restricciones, permitiéndose incumplir un número mínimo de ellas para optimizar la función objetivo, de esta manera surge un compromiso entre el número de restricciones incumplidas y el valor de la función objetivo. Este tipo de métodos emplean diversas técnicas para lograr obtener una partición atendiendo a las restricciones:

- Modificar la función objetivo de manera que incluya una penalización por incumplir restricciones.

- Agrupar con información adicional obtenida de una distribución condicional en un espacio auxiliar.
- Forzar el cumplimiento de todas las restricciones modificando la regla de asignación del método.
- Inicializando los clusters e base a restricciones inferidas del conjunto de instancias etiquetadas.

La figura 3.2 muestra un conjunto de datos junto a sus restricciones asociadas, la 3.3 propone un posible agrupamiento que satisface todas las restricciones.

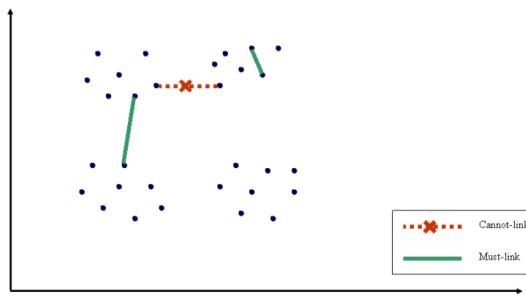


Figura 3.2: Restricciones sobre un conjunto de datos [1]

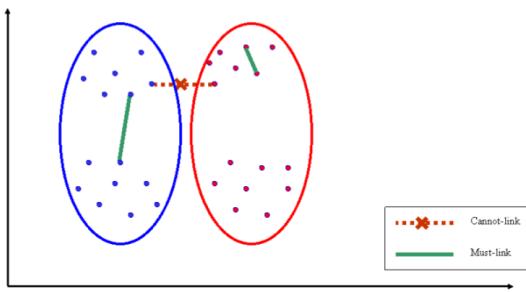


Figura 3.3: Clustering que satisface todas las restricciones [1]

### *Métodos basados en distancia*

En las aproximaciones basadas en distancias, se emplean métodos de clustering clásicos que hagan uso de una medida de distancia, de forma que dicha medida se modifica para que tenga en consideración las restricciones. En este contexto, satisfacer las restricciones significa que las instancias relacionadas con restricciones Must-Link (ML) se sitúan juntas en el espacio, y las relacionadas mediante Cannot-Link (CL) se encuentran separadas.

La figura 3.5 muestra un posible agrupamiento basado en una métrica aprendida a partir de las restricciones especificadas en la figura 3.4. Cabe destacar que en la figura 3.5 el espacio en el que se encuentran los datos ha sido comprimido en el eje vertical y ensanchado en el eje horizontal para ajustarlo a la métrica de distancia aprendida.

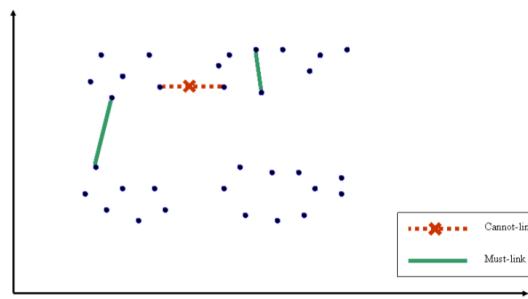


Figura 3.4: Restricciones sobre un conjunto de datos [1]

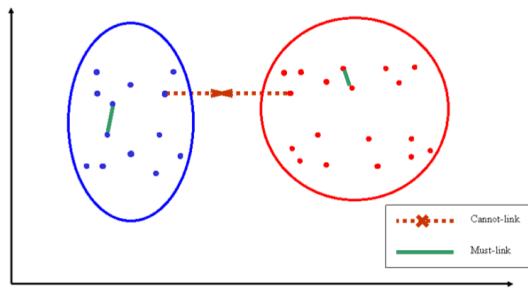


Figura 3.5: Clustering basado en métrica aprendida en base a las restricciones [1]

#### APLICACIONES DEL CLUSTERING CON RESTRICCIONES

Este epígrafe muestra algunos casos de aplicación en los que el clustering con restricciones ha resultado ser una herramienta más útil que el clustering no supervisado. Para cada caso analizaremos como se obtuvieron las restricciones y como estas mejoran los resultados en el clustering resultante.

##### *Aplicaciones en análisis de imágenes*

La figura 3.6 muestra un extracto del conjunto de datos de caras de Carnegie Mellon University (CMU), en el que la tarea es agrupar caras en base a diferentes criterios. En este caso, el objetivo es agrupar las caras según su orientación, es decir, las caras con la misma orientación deberán estar en el mismo cluster.

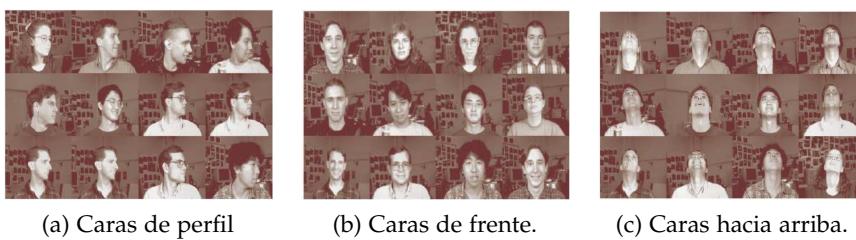


Figura 3.6: Caras de la base de datos de CMU [1]

El método empleado para extraer las restricciones es uno de los más populares en la literatura: establecer el número de clusters igual al número de clases en la base de datos, y generar las restricciones a partir de un subconjunto de instancias etiquetadas, esto es, si dos instancias tiene diferentes etiquetas, establecer una restriccion Cannot-Link (CL) entre ellas, en caso contrario una de tipo Must-Link (ML). De esta forma, entre las imágenes mostradas en la figura 3.7 se establecen restricciones Cannot-Link (CL), ya que, aunque pertenecen a la misma persona, no presentan la misma orientación.



Figura 3.7: Restricciones de tipo CL entre caras de la misma persona [1]

En la figura 3.8 se muestra otro conjunto de datos de imágenes sobre el que se aplican técnicas de clustering con restricciones. En este caso, la tarea es realizar reconocimiento de objetos para incorporar el método al sistema de navegación del robot Aibo. Para ello se emplean restricciones de distancia de tipo  $\delta$  y  $\epsilon$  como las descritas en la figura 3.1, de esta manera se consiguen clusters bien diferenciados y por tanto útiles para las tareas de búsqueda de caminos que el robot realiza durante la navegación.

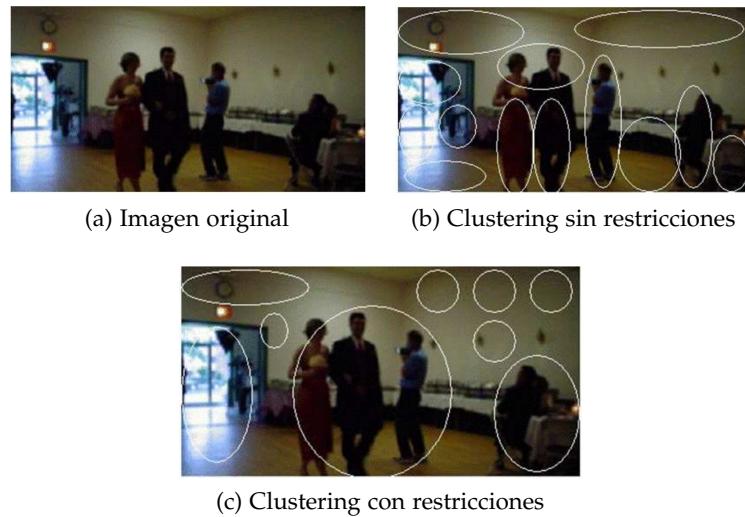


Figura 3.8: Método de clustering empleado en el sistema de navegación del robot Aibo [1]

### *Aplicaciones en análisis de videos*

Las bases de datos de video son uno de los ejemplos en los que las restricciones pueden ser generadas directamente desde el dominio de datos, especialmente disponiendo de datos espacio-temporales sobre el vídeo. En datos temporalmente sucesivos es posible establecer restricciones de tipo Must-Link (ML) entre grupos de píxeles de fotogramas (*frames*) cercanos en el tiempo. Esto es especialmente útil cuando la tarea es implementar reconocimiento de objetos basado en clustering y segmentación. También es posible añadir restricciones Cannot-Link (CL) a clusters localizados en el mismo fotograma, ya que existe una baja probabilidad de que estén asociados al mismo objeto. De hecho, en el dominio asociado a problemas de análisis de vídeo existen gran variedad métodos de extracción de restricciones, la figura 3.9 muestra algunos ejemplos.

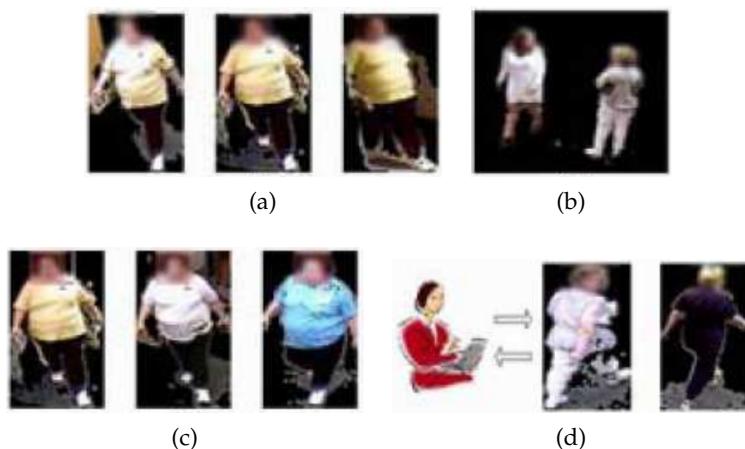


Figura 3.9: Diferentes tipos de restricciones en datos de video [1]

En la figura 3.9, la imagen (a) corresponde a restricciones extraídas del seguimiento de una persona durante un periodo de tiempo, la (b) corresponde a restricciones espaciales que asocian dos objetos localizados en el mismo fotograma, la imagen (c) corresponde a restricciones obtenidas mediante reconocimiento facial y la (d) a las proporcionadas por el usuario.

Disponiendo de tantos métodos para extraer restricciones, una cuestión que cabe plantearse en este contexto es, ¿qué sucede si se establecen demasiadas restricciones? ¿Hace esto que el problema esté sobre-restringido? En epígrafes posteriores abordaremos estas cuestiones.

### *Aplicaciones en genética*

En clustering de genes basado en microarrays, los genes vienen representados por su perfil de expresión en diferentes experimentos y agrupados empleando diferentes algoritmos, en este caso algoritmos

de clustering con restricciones. La figura 3.10 muestra un ejemplo, en este caso las restricciones de tipo Must-Link (ML) se establecen entre genes en base a los datos de co-ocurrencia almacenados en la base de datos de interacciones de proteínas, que contiene información sobre qué genes (y sus proteínas asociadas) están asociados a los mismos procesos celulares.

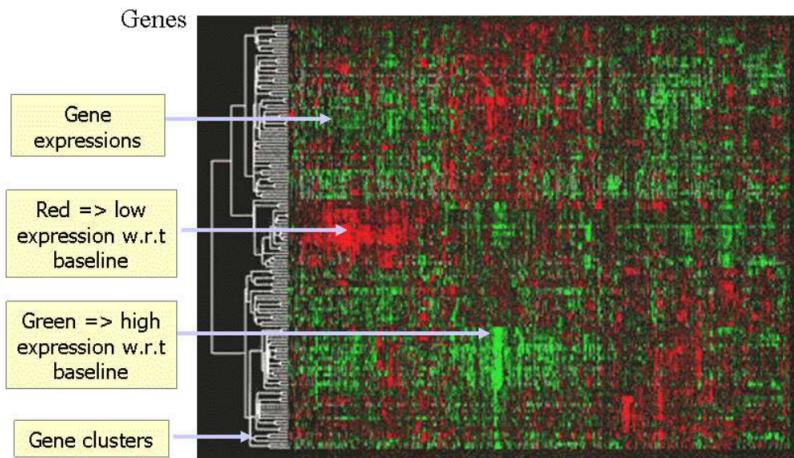


Figura 3.10: Clustering de genes basado en microarrays [1]

*Aplicaciones en análisis de textos*

## BIBLIOGRAFÍA

---

- [1] Ian Davidson y Sugato Basu. «A Survey of Clustering with Instance Level Constraints». En: *ACM Transactions on Knowledge Discovery from Data* (2007), págs. 1-41.
- [2] S. Brian Everitt, Sabine Landau, Morven Leese y Stahl Daniel. *Custer Analysis*. Berlin, Germany: WILEY, 2011.