Universidad Miguel Hernández

Grado en Estadística Empresarial

Resolución de análisis clúster mediante problemas de optimización combinatoria

Autor: José Navajas Bañón

Tutora: Mercedes Landete Ruiz

Curso 2020/2021

Índice General

Lista de tablas

Resumen

En la actualidad, la recogida de datos se ha convertido en un proceso vehicular en el avance de la sociedad y la toma de decisiones. Diversos son los ámbitos de acción en los que el análisis de los datos es imprescindible, entre otros, los datos en el espacio de la psicología en el que a través de ellos se pueden recoger patrones de comportamiento, en el mundo empresarial en general con tan distintas aplicaciones como en la logística, como en la asignación de tareas a trabajadores, tiempos de espera... Así como en instituciones gubernamentales, en la gestión del territorio, ampliación de servicios y una larga enumeración en distintas áreas.

Por ello, se ha considerado pertinente realizar una introducción acerca del análisis clúster que, debido a la sencillez de su funcionamiento, existe una constante inquietud en su aplicación en múltiples ramas de estudio. El desarrollo computacional, como nuevas técnicas de resolución en menor tiempo y esfuerzo hace que no se necesiten una gran cantidad de recursos para conseguir realizar este proceso de clasificación. Así, se expone un análisis sobre la teoría del análisis clúster, tipología de clasificación, naturaleza de los datos sobre los que aplicar...

En un segundo acto, se propone una revisión acerca de la literatura del problema *p-median* así como métodos heurísticos, utilizado en la mayoría de casos, para la ubicación óptima de instalaciones. Debido a este gran crecimiento en el estudio y recolección de datos, ha hecho que uno de los campos en los que mayor interés y crecimiento han desarrollado es el del sector logístico. Uno de los problemas que más inquietud tiene, es el de la ubicación de almacenes con los que poder abastecer a los clientes, ubicación potencial de comercios para el consumidor y satisfacer las necesidades de estos, de igual manera, instituciones gubernamentales plantean este hecho con el fin de proporcionar servicios a toda la población aplicando eficientemente los recursos disponibles.

Análisis cluster

En este apartado del trabajo se realiza una exposición acerca de la técnica de agrupación clúster de datos, así como distintos tipos de análisis en función de la tipología de las variables, distintas técnicas posibles, etc...

El análisis clúster es una técnica utilizada en gran cantidad de campos de estudio, como pueden ser, entre muchos otros en economía, sociología, psicología, en biología, teniendo como punto en común estos campos la clasificación entre sujetos y objetos según la naturaleza de los datos. A menudo, la información conocida sobre las categorías es nula o muy baja, siendo el objetivo primario descubrir la estructura de los datos. El análisis clúster, se trata de una técnica multivariante mediante la cual se pretende clasificar una serie de objetos o variables, generando grupos/conglomerados, a priori desconocidos. Se intenta garantizar la heterogeneidad entre los distintos grupos formados, mientras que los objetos de cada grupo (clúster) sean lo más homogéneos entre sí.

Dichos conglomerados se generan siguiendo algún criterio de homogeneidad como: medidas de distancia o similitud entre las observaciones, reglas de asociación, funciones de pérdida... El análisis clúster distingue dos métodos mediante los cuales, en función de los datos disponibles, realiza el estudio diferenciando: métodos jerárquicos y métodos no jerárquicos.

En esencia, la diferencia entre ambos radica en que el método no jerárquico clasifica los objetos en K grupos/clústers determinados con anterioridad, es decir se comienza el análisis conociendo el número de grupos, siendo estos disjuntos entre sí, mientras que en los métodos jerárquicos no es necesario K grupos/clústers previos, ya que a través del proceso iterativo se generan sucesiones ordenadas (jerarquías) de clústers.

Planteamiento general del problema

El planteamiento del análisis clúster, considerando una matriz X, la cual proporciona los valores de las variables *(p)* para cada uno de los individuos *(n)*:

$$\mathrm{X} = \begin{pmatrix} x_{11} & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{ij} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

La i-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores de la j-ésima variable para el conjunto de individuos de la muestra. Se tiene por objetivo agrupar los objetos representados en las filas de la matriz X en distintos grupos/clústers.

Se trata de un proceso mayoritariamente descriptivo, ya que en su proceso no suele utilizar modelos estadísticos para la agrupación de los objetos. Estructurando las etapas del análisis se tiene:

1. En primer lugar, partiendo de una matriz X (nxp), seleccionar las variables relevantes para el tipo de clasificación que se pretende.
2. Seguidamente, el análisis requiere algún sistema de asociación, como la medición de proximidad entre los objetos, generalmente en términos de distancias, mientras que, para la agrupación de variables, la medida utilizada suele venir dada por el coeficiente de correlación.
3. El tercer paso, corresponde con la elección del algoritmo de agrupación, diferenciando entre los métodos jerárquicos y no jerárquicos.
4. Finalmente, tras seleccionar el método de clasificación se procede a la interpretación y validación de los resultados obtenidos. Para ello, distintos autores proponen el uso de técnicas multivariantes entre otras, como puede ser: Análisis Factorial, Análisis Discriminante, ANOVA, MANOVA...

Criterios de similitud

Partiendo de la matriz X (nxp), existen distintas métricas que cuantifican la relación entre cada par de objetos. Las métricas de distancia o similitud entre cada par de objetos dan origen a una matriz de distancias, en la que cada componente representa el valor de la métrica utilizada entre los individuos i y j.

Se tienen las medidas de proximidad, similitud o semejanza que miden el grado de igualdad entre un par de objetos, por lo que, cuando esta medida sea grande, mayor homogeneidad entre los objetos. En caso contrario, las medidas de distancia o disimilaridad, en las cuales, cuanto mayor sea el valor, la similitud entre el par es menor, por lo que más diferentes serán lo que conlleva a una probabilidad menor de pertenencia al mismo grupo/clúster.

La selección de la medida de similitud es uno de los motivos que mayor problema puede ocasionar sobre los resultados, por lo que su elección depende de muchos factores a tener en cuenta (tipología de las variables, etc.).

Se distinguen entre medidas de asociación para variables e individuos, aunque dichas medidas pueden ser utilizados en ambos problemas, solucionándolo a través de la trasposición de la matriz de los datos originales. La tipología de las variables posibles en el estudio puede ser:

- De intervalo o razón: todas las variables son cuantitativas.
- Frecuencias: variables categóricas
- Binarios: se trata de una matriz de objetos en el que las variables son binarias, indicando 1 como presencia de la característica y 0 en caso de ausencia.

Medidas de asociación entre individuos

En el siguiente apartado se presenta algunas de las medidas de asociación entre individuos más repetidas en el análisis clúster, distinguiendo entre ellas según la tipología de la variable.

Distancia para variables cuantitativas

A continuación, se presentan diversas medidas númericas expresando el grado de asociación entre variables cuantitativas, es decir, aquellas que dan como resultado un valor numérico.

Distancia euclídea

Esta métrica es la más utilizada en el análisis clúster, la cual mide la distancia entre el objeto i y el objeto j, dada por:

$$d(x_i, x_j) = ||x_i - x_j||_2 = \sqrt{(x_i - x_j)'(x_i - x_j)} = \sqrt{\sum_{l=1}^{n}(x_{il} - x_{jl})^2}$$

Esta distancia presenta varios inconvenientes a la hora de su aplicación: se trata de una medida sensible a cambios de escala y considera que las variables utilizadas en su cálculo son independientes.

Por lo que, la escala en la que estén recogidas las variables influirá de manera que, aquellas que contengan valores altos contribuirán en mayor medida al resultado, siendo una posible solución la tipificación de las variables.

Mientras que la presunción de independencia entra las variables provoca que para aquellas que estén correlacionadas aumente la disimilaridad entre los individuos. Para ello es posible corregirlo mediante la sustitución de las variables originales por componentes principales, siendo estas incorreladas entre sí. Otra posible solución es la ponderación de las variables con pesos inversamente proporcional a las correlaciones.

Distancia de Minkowsky

La medida de distancia de Minkowsky se define:

$$d(x_i, x_j) = \sqrt[q]{\sum_{l=1}^{p}(X_{il} - X_{jl})^q}, \quad q \in N$$

Esta medida tiene el inconveniente que no se muestra invariante a cambios de escala. Por otra parte, según los valores asignados al parámetro p, se tienen distintas distancias:

Distancia de Manhattan

Distancia de Manhattan ($p = 1$), definida como la suma del valor absoluto de la diferencia de cada par de objetos, la cual no se ve afectada por valores extremos (outliers).

$$d(x_i, x_j) = \sum_{l=1}^{p}|X_{il} - X_{jl}|$$

Distancia máxima o de Chebyshev

Distancia máxima ($p = \infty$), expresa la magnitud absoluta de la diferencia entre coordenadas de un par de objetos.

$$d(x_i, x_j) = max_l|x_{il} - x_{jl}|$$

Correlación entre individuos

El coeficiente de correlación puede ser utilizado como una métrica de asociación entre individuos.

$$r_{ij} = \frac{\sum_{l=1}^{n}(x_{il} - \overline{x}_i)(x_{ij} - \overline{x}_j)}{s_i s_j}$$

Donde se ha definido,

$$\overline{x}_h = \frac{1}{n}\sum_{l=1}^{n} x_{hl} \quad h = i, j, \qquad \textit{media de cada individuo}$$

$$s_h^2 = \sum_{l=1}^{n}(x_{hl} - \overline{x}_h)^2 \quad h = i, j, \qquad \textit{desviación cuadrática de cada individuo}$$

La diferencia entre las unidades de medida en el vector de observaciones de cada individuo dificulta la comparación de las medias y varianzas. Así, los cambios de escala provocan que el coeficiente de correlación varíe, y pueda dar lugar a resultados equivocados.

Distancia para variables categóricas

Distancia basada en el estadístico Chi-Cuadrado

El estadístico Chi-Cuadrado (χ^2), normalmente utilizado para determinar si un modelo estadístico ajusta los datos correctamente. En este caso, se puede utilizar para medir diferencias (distancia) entre dos variables categóricas.

Para este fin, se recogen los datos en las llamadas tablas de contingencia, las cuales contienen el total de valores observados en las distintas categorías de las variables correspondiente.

Notemos:

- o_{ij} = valor (frecuencia) observado en la posición i, j.
- e_{ij} = valor (frecuencia) esperado bajo la hipótesis de independencia.

A continuación, se muestra la estructura necesaria para la medición de las diferencias entre un par de variables categóricas

$VarA$ \ $VarB$	1	$\dots$	j	$\dots$	q	
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	n_{il}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i.}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p.}$
	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.q}$	$n_{..}$

Se define el estadístico χ^2 como:

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

donde p y q corresponden al número de categorías que toman las variables en estudio.

Distancia para variables binarias

En este apartado, se explican algunas métricas para medir la relación de las variables binarias, las cuales suelen tomar valor *{0,1}*, siendo 1 cuando la variable tiene la presencia de esa característica, y 0 en caso de ausencia.

Ind. I \ Ind. J	1 0	Totales
1	a b	a + b
0	c d	c + d
Totales	a+c b+d	a+b+c+d

Se pueden definir los distintos valores de la tabla como:

- a: número de variables con respuesta 1 en ambos individuos *(i,j)*.
- b: número de variables con presencia i, pero 0 en el individuo j.
- c: variable con valor 1 en el individuo j, pero 0 en el individuo i.
- d: número de variables con respuesta 0 en ambos individuos *(i,j)*.

Alguna de las medidas más populares para calcular la distancia en variables binarias son:

- *Lance and Williams*: medida también conocida como coeficiente no métrico de Bray-Curtis.

  $$d = \frac{(b + c)}{2a + b + c}$$

  Los posibles valores que toma se encuentran en el rango [0-1].
- *Distancia euclídea*:

  $$d = \sqrt{(b + c)}$$

Medidas de asociación para variables

En este apartado se explican algunas de las medidas numéricas a partir de las cuales, se consigue agrupar variables en un análisis clúster. Entre las que se pueden encontrar:

- Ochiai.
- Medida ϕ.
- Medida de Russell and Rao.
- Medida de Jaccard.
- Medida de Dice.
- Medida de Roger-Tanimoto.
- etc.

Coeficiente de correlación

Considerando las variables X_i y X_j centradas respecto a sus medias, se tiene para la muestra de m individuos:

$$\hat{x}_i = (x_{li} - \bar{x}_i, ..., x_{mi} - \bar{x}_i)$$
$$\hat{x}_j = (x_{lj} - \bar{x}_j, ..., x_{mj} - \bar{x}_j)$$

La correlación muestral entre x_i y x_j se define:

$$r = \frac{Cov(x_i, x_j)}{(Var(x_i)Var(x_j))^2} = \frac{\sum_{l=1}^{m}(x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j)}{(\sum_{l=1}^{m}(x_{li} - \bar{x}_i)^2 \sum_{l=1}^{m}(x_{lj} - \bar{x}_j)^2)^{\frac{1}{2}}}$$

Se tiene que la correlación es invariante ante transformaciones lineales, a excepción de cambios de signo, el coeficiente de correlación usa los datos centrados empleando desviaciones respecto a la media.

Medidas de asociación para datos dicotómicos

Se entiende por variable dicotómica, aquella que usualmente toma dos valores identificando la ausencia como valor 0 y la presencia de la característica como 1.

La relación entre dos variables dicotómicas genera una tabla de contingencia $2x2$, tal que:

$X_i \setminus X_j$	1	0	Totales
1	a	b	a + b
0	c	d	c + d
Totales	a+c	b+d	a+b+c+d

Según la tabla se considera:

- a: representa el número de individuos que toman el valor 1 en cada variable al mismo tiempo.
- b: número de individuos que toman el valor 1 en la variable X_i y 0 en la X_j.
- c: número de individuos que toman el valor 0 en la variable X_i y 1 en la X_j.
- d: representa el número de individuos que toman el valor 0 en cada variable simultáneamente.

Medidas basadas en coincidencias

Una forma de medir la similaridad en variables binarias es el de contar las veces que ocurre un suceso en dos variables al mismo tiempo. Así, dos variables son coincidentes simultáneamente cuanto mayor sea el número de coincidencias.

$$\frac{a+d}{a+b+c+d}$$

Medida de Russell and Rao

Esta medida recoge la probabilidad de que, al azar, un individuo tenga el valor 1 en ambas variables. Asímismo, la medida proporciona igual peso a las coincidencias y no coincidencias.

$$\frac{a}{a+b+c+d}$$

Medida de Jaccard

El coeficiente de *Jaccard* no tiene en cuenta cuando ambas variables toman el valor 0, debido a que en algunas situaciones las variables binarias son asimétricas, es decir, resulta más informativo o relevante la presencia del atributo.

$$\frac{a}{a+b+c}$$

Métodos de clasificación de los métodos clúster

Una vez definidas algunas de las principales medidas de asociación tanto para variables como para individuos, se procede a la explicación de los distintos métodos de clasificación.

Para ello, en primer lugar, se destacan dos grandes grupos de análisis: métodos jerárquicos y métodos no jerárquicos.

Método clúster jerárquico

El análisis clúster jerárquico tiene por objetivo agrupar en un grupo nuevo o dividir uno ya existente mediante un proceso iterativo jerárquico, buscando para ello la maximización o minimización de alguna medida de distancia o similitud.

Este método se diferencia a su vez, en función a la manera de formar los clúster entre, aglomerativos o divisivos.

Métodos jerárquicos aglomerativos

Los métodos jerárquicos algomerativos parten con un número de clúster igual al número de individuos existentes. A partir de este punto, se van agrupando en sucesivos pasos, aquellos con características similares hasta conseguir agrupar todas las observaciones.

Existen distintas formas de generar los clúster en los distintos niveles del método jerárquico. No puede establecer de antemano la mejor estrategia para este fin, por lo que se debe realizar el estudio con distintas estrategias de agrupación.

Se tiene:

- **Estrategia de la distancia mínima o similitud máxima**: se considera que la distancia o similitud entre dos clúster viene dada, respectivamente, por la mínima distancia (o máxima similitud) entre sus componentes.
- **Estrategia de la distancia máxima o similitud mínima**: la distancia o similitud entre clúster viene dada, respectivamente, por la máxima distancia (o mínima similitud) entre sus componentes.
- **Estrategia de la distancia, o similitud, promedio no ponderado**: se considera que la distancia o similitud, entre dos clúster se obtiene como la media aritmética entre la distancia o similitud, de las componentes de dichos clúster.
- **Métodos basados en el centroide**: la semejanza entre dos clúster viene dada por lo semejante entre sus centroides.
- **Método de Ward**: en cada iteración se unen los dos clúster cuyo incremento del valor total de la suma de cuadrados de las diferencias sea menor, dentro de cada clúster.

Los métodos clúster jerárquicos aglomerativos son:

- **Aglomerativos**.
  - Método de Linkage Simple, o Vecino más cercano.
  - Método de Linkage Completo, o Vecino más alejado.
  - Método Promedio entre grupos.
  - Método del Centroide.
  - Método de la Mediana.
  - Método de Ward.

Métodos jerárquicos divisivos

Por otra parte, se encuentran los métodos jerárquicos divisivos. A diferencia de los métodos aglomerativos, el punto inicial radica en todas las observaciones agrupadas en un único clúster, a partir del cual, se van formando tras sucesivas particiones grupos más pequeños. El límite divisivo llega cuando cada observación pertenece a un grupo diferente.

Los métodos clúster jerárquicos divisivos son:

- **Divisivos**.
  - Método de Linkage Simple, o Vecino más cercano.
  - Método de Linkage Completo, o Vecino más alejado.
  - Método Promedio entre grupos.
  - Método del Centroide.
  - Método de la Mediana.
  - Método de Ward.
  - Método de análisis de asociación.

Método clúster no jerárquico

Por su parte, el análisis clúster no jerárquico, conocido como partitivo o de optimización, a diferencia del análisis clúster jerárquico, el número de k clúster en los que agrupar las observaciones vienen especificados a priori. Se pretende asignar cada objeto al grupo/clúster que presenta mayor coincidencia en sus características, para ello optimiza algún criterio de distancia, función de coste, etc.

El objetivo por lo tanto, es el de conseguir grupos homogéneos de objetos dentro de él, y que la varianza entre los grupos sea grande. A diferencia del análisis clúster jerárquico, los clúster formados no guardan relación o jerarquías entre sí, si no que se encuentran claramente separados entre ellos. Tras establecer los k clúster, intercambiando de clúster los objetos en cada iteración hasta conseguir la mejor asignación posible. El diseño de este tipo de método de análisis no permite agrupar variables por si solos, a diferencia de los análisis clúster jerárquicos.

Entre los métodos no jerárquicos se puede diferenciar entre:

- **Métodos de reasignación**
  - Método de K-Medias.
  - Método K-Medoids o PAM.

  - Método de Forgy.
  - Método de Nubes Dinámicas.
  - Quick-Cluster análisis.
- **Métodos de búsqueda de densidad**
  - Análisis modal de Wishart.
  - Método Taxmap.
  - Método de Fortin.
- **Métodos directos**
  - Método de Block-Clustering.
- **Métodos de reducción de dimensiones**
  - Análisis factorial tipo Q

Métodos de reasignación

Este tipo de métodos permite la reasignación de los individuos en cada paso, de su clúster asignado a otro, si con ello se optimiza el criterio de selección o de parada, el individuo es reasignado a ese nuevo grupo, finalizando el proceso una vez que no se consiga optimizar el criterio establecido.

Método K-Means

McQueen, (1972), desarrolla el método *K-Mean* asignando cada individuo al clúster cuyo centroide sea el más próximo. Se trata de uno de los métodos de análisis clúster más actualizados en la actualidad, ya que en la mayoría de casos, se trata de un método rápido, sencillo y con una correcta inicialización (nº de clúster), genera resultados buenos.

Una de las premisas en este método es la obligatoriedad de que todas las variables del estudio sean cuantitativas. En cuanto a la medida de distancia que utiliza el algoritmo entre los objetos medidos, es la distancia euclídea. Debido a la sensibilidad de esta medida a cambios de escala, se recomienda la estandarización de las observaciones para que el resultado no se vea afectado.

Una de las características más relevantes del método, es la reasignación continua de las observaciones a distintos clúster, ya que en cada asignación los centroides se recalculan. El objetivo del algoritmo es el de minimizar la suma de cuadrados dentro del clúster. El algoritmo propuesto por McQueen sigue:

1. Se selecciona a priori, k centroides siendo k a su vez el número de clúster.
2. Asignación del resto de individuos al clúster con el centroide más cercano al individuo, en función a la distancia euclídea entre ambos. Una vez asignado el individuo, se recalculan los centroides de cada clúster.
3. El algoritmo finaliza cuando se llega a un criterio de parada establecido.

El algoritmo resuelve un problema de optimización, cuya función objetivo consiste en minimizar la suma de las distancias cuadráticas de cada objeto al centro (centroide) del clúster, donde $S = 1, ...,k$ es el conjunto de datos cuyos elementos son lo objetos x_j.

$$f.o.min = \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2$$

Las ventajas de utilizar el algoritmo *k-means* reside en la rapidez y la sencillez, donde un posible problema viene dado por la obligatoriedad de suponer a priori un número k de conglomerados, pudiendo generar agrupaciones poco óptimas si no se selecciona el número k correcto. Una de las desventajas de este análisis es la influencia de valores atípicos

Método K-Medoids

El algoritmo *k-medoids (PAM)*, es un método de reasignación en el cual considera como objetos representativos en cada uno de los clúster, aquel más centrado en el conglomerado en lugar del centroide. Esta casuística hace que el algoritmo sea robusto a los valores atípicos, sin que esto pueda distorsionar el resultado.

Por otra parte, *k-medoids* funciona de forma ineficiente, debido al tiempo de computación, si el tamaño de los datos es muy grande. Se tiene n objetos con p variables que deben ser agrupados en $(k < n)$ clúster, con k definido a priori. Se utiliza la distancia euclídea como medida de disimilitud. El algoritmo se compone de las siguientes etapas:

- Selección de medoides.
  - Calcular la distancia entre cada par de objetos (distancia euclídea)
  - Calcular v_j como: $v_j = \sum_{i=1}^{n} \frac{d_i j}{\sum_{l=1}^{n} d_{il}}$, $i = 1, \ldots, n; j = 1, \ldots, n$
  - Seleccionar los k objetos que tengan los primeros k valores más pequeños, como centros de conglomerado iniciales.
- Actualiza los centros de conglomerados por aquellos que minimicen la distancia total a otros objetos del clúster.
- Asigna los objetos al centro más cercano.
- Recalcular la distancia de totos los objetos a sus medoides, y si la suma es igual a la anterior, se detiene el algoritmo, en caso contrario, se vuelve al Paso 2.

Métodos de búsqueda de densidad

Método mediante el cual, los clúster se forman buscando zonas con una gran concentración de individuos, diferenciando los que producen un resultado tipológico o probabilístico.

Métodos directos

Este método permite clasificar simultáneamente individuos y variables, cuyo algoritmo más representativo es el *Block-Clustering*.

Métodos de reducción de dimensiones

Este método está basado en el descubrimiento de grupos con objetos semejantes, esperando que tengan algún tipo de propiedad en común. La matriz de entrada con la que se inicia el algoritmo es la de correlaciones entre los individuos, siendo el problema que cada individuo puede pertenecer a ambos grupos y presentar solapamiento, por lo que su interpretación se vuelve compleja.

Validación de los resultados

En este apartado, se describen ciertas medidas con las que determinar la validez de los resultados obtenidos por el análisis de clustering, puediendo ser técnicas multivariantes, análisis de la varianza, etc.

Esta situación ha generado que estos aspectos dependan del juicio del investigador, ya que no existen criterios unificados entre investigadores sobre una mejor solución en la validación de los resultados obtenidos.

En el análisis clúster, no todos los grupos/clúster formados tienen que ser significativos, por ello, para una correcta interpretación, se puede realizar:

- Realizar un análisis de la varianza (ANOVA), ya que en este método estudia la variabilidad existente dentro de cada uno de los grupos, como la variabilidad entre los distintos grupos.
- Validación cruzada, es decir, repetir el análisis clúster sobre distintas muestras y comprobar que el resultado ofrecido es similar al que nos proporciona el análisis sobre todos los datos.
- Aplicación de índices como el Índice de silueta [1], que estima el número óptimo de clúster existentes en los datos, el índice de validación externa, comúnmente usando la etiqueta de los datos con el fin de validar la validez de los grupos como el Accuracy (ACC) [14].
- Análisis Discriminante, aplicándolo sobre cada clúster formado, teniendo como variable dependiente el clúster de pertenencia y el resto de variables como predictoras observando el grado de discriminancia de cada una.[16]

Revisión bibliográfica de artículos científicos

Se procede a la recolección de artículos de interés sobre el estudio del problema *p-median*, desde sus variantes, como el problema de la *p-median capacitada*, hasta la resolución del problema mediante técnicas heurísticas y metaheurísticas.

Así mismo, como para distintos usos de aplicación, como es la asignación de objetos a grupos hasta la reducción de variables de los datos disponibles.

The p-median model as a tool for clustering psychologial data

El artículo publicado en 2010, por los autores *Hans-Friedrick Köhn*, *Douglas Steinley* y *Michael J. Brusco*, *'The p-median Model as a Tool for Clustering Psychological Data'*, profundiza en la aplicación del modelo *p-median* para la resolución de agrupaciones en los datos.

Esencialmente, las agrupaciones basadas en clúster pueden clasificarse como una serie de métodos que buscan identificar grupos en la estructura de los datos, así como observar diferencias entre los grupos de la muestra.

En el ámbito de la psicología, los modelos de agrupamiento resultan accesibles y de una fácil interpretación debido a su lógica de actuación, similar a procesos cognitivos del ser humano, tal como recogieron *Murphy, 2002; Ross, Taylor, Middleton y Nokes, 2008* en relación con la teoría ejemplar (forma en la que el ser humano categoriza objetos). Conceptualizando ambos como: dado un conjunto de datos, se seleccionan determinados objetos como centro de los conglomerados, asignando el resto a un conglomerado cercano, es decir más símil, en función a distintos criterios.

Debido a la asignación de objetos reales como centro de los grupos en el problema *p-median*, que a diferencia del método *k-Means* cuyo centroide no es un objeto de la base en sí, genera mayor aceptación entre los investigadores de distintas áreas. La aplicación del problema *p-median*, al igual que *k-Means*, genera un conjunto de grupos disjuntos y no superpuestos, teniendo como ventaja la aplicación a mayor variedad de datos, así como su robustez ante datos atípicos.

Con el fin de una mejor interpretación del problema *p-median*, se remarcan una serie de conceptos teóricos. Primeramente, se introduce el concepto de proximidad como cualquier medida numérica que relaciona distintos pares de elementos. Estas proximidades son recopiladas normalmente en matrices, conteniendo los objetos y el valor de su *'proximidad'*.

Por otro lado, la función de pérdida (Loss Function), recoge la diferencia numérica entre los valores reales y predichos, siendo útil para medir la bondad del modelo.

El último concepto descrito es el de óptimo global/local. Este concepto está comprendido dentro de una función de pérdida, donde la diferencia radica en que un óptimo global indica que no existe una solución mejor en el conjunto, mientras que un óptimo local ofrece la mejor solución dentro de un subgrupo.

Una vez introducidos los conceptos anteriores y modelos de agrupamiento, se describe con mayor detalle la *p-median* como un problema de optimización discreta, en el que es asignado a la clase C_k o $C_{k'}$ (donde $1 \leq k,\ k' \leq K$, con K igual al número total de clases). El conjunto de soluciones es finito, existiendo siempre un óptimo global, sin embargo, debido a la enumeración de todas las agrupaciones posibles, siendo computacionalmente muy difícil. Esta problemática se intenta resolver a partir de estrategias de enumeración parcial como *Branch-and-Bound (Brusco y Stahl, 2005)*[18] o mediante programación dinámica *(Hubert, Arabie y Meulman, 2001)*[17].

El modelo de agrupación *p-median* se origina con el fin de optimizar el problema de ubicación de instalaciones. Se modeliza a partir del siguiente problema de optimización: dado un conjunto de N elementos, seleccionando p grupos (medianas), asignando los $N - p$ objetos restantes a las p mediandas minimizando la función de la suma total de diferencias entre los objetos y las medianas.

En su forma general, puede utilizarse una matriz rectangular de proximidades entre dos conjuntos distintos de entidades, siendo habitual en ciencias sociales, matrices cuadradas de distancias entre pares de elementos de un mismo conjunto de datos.

La formulación específica de la agrupación del problema *p-median*, es:

$$\min_{x} \; f(X) = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} x_{ij}$$

s.a.:

$$\sum_{j=1}^{N} x_{jj} = p, \tag{1}$$

$$x_{ij} \leq x_{jj} \qquad \forall i, j, \tag{2}$$

$$\sum_{j=1}^{N} x_{ij} = 1 \qquad \forall i. \tag{3}$$

donde:

$D_{NxN} = d_{ij}$, distancia entre i y j

$$x_{ij} = \begin{cases} 1, \text{si un objeto } i \text{ se asigna a una mediana j} \\ 0, \text{en caso contrario} \end{cases}$$

$$x_{jj} = \begin{cases} 1, \text{si un objeto } j \text{ se selecciona como mediana} \\ 0, \text{en caso contrario} \end{cases}$$

Por lo que, x_{ij} denotan las variables de decisión, mientras que x_{jj} hace referencia a los elementos de la diagonal de la matriz X.

La formulación del problema establece que, si una variable de decisión (x_{ij}) se establece como 1, el valor de la celda en D (matriz de distancias) se evalúa en la función objetivo. La restricción (1), obliga al problema a que la suma de las variables x_{jj} en la diagonal sea igual a p (nº total de medianas). En segundo lugar, la restricción (2) asegura que x_{ij} solo puede ser asignada a una variable x_{jj} seleccionada como mediana. Por último, en (3) se establece que x_{ij} debe ser asignada a alguna mediana activa (x_{jj}).

Respecto a la búsqueda de un óptimo global para el problema *p-median*, se propone un método de tres etapas *(Brusco y Köhn, 2008)*[15]: (1) Sustitución de Vértices (VS);(2) relajación lagrangiana (LR)/optimización de subgradiente; (3) Branch-and-Bround.

El método de *Sustitución de Vértices (VS)*, realiza un proceso iterativo en el que, partiendo de posibles candidatos a mediana realiza sustituciones con los vértices no seleccionados como candidatos, finalizando cuando una sustitución no produzca mejora, llegando así a óptimos locales. Esta etapa, al tratarse el problema *p-median* de minimización, proporciona un límite superior a la solución óptima.

La segunda etapa de optimización, *relajación Lagrangiana (LR)*, consiste en eliminar aquellas restricciones más difíciles y añadirlas como una penalización en la función objetivo mediante ponderaciones (Multiplicadores de Lagrange). A través del algoritmo de optimización de subgradiente, se resuelven los problemas de la relajación lagrangiana (LR), siempre que se disponga de un límite superior (VS), a partir de la solución proporcionada por la LR.

Si la solución no es globalmente óptima, se inicia la búsqueda del óptimo mediante Branch-and-Bound. Su metodología trata de dividir un problema en varios subproblemas, enfrentando cada uno a cotas inferiores (VS/LR). Finalmente, se examinan los subprocesos sobre los que acotar la búsqueda hacia el óptimo, descartando el resto de ramificaciones.

Igual que en muchos métodos de agrupamiento, se establece los p grupos a priori. Con el fin de alcanzar una solución óptima, *(Rousseeuw, 1987)*[2], definió el índice de silueta para determinar el número de grupos p:

$$SI_k = \sum_{i=1}^{N} \frac{b(i) - a(i)}{max\{a(i), b(i)\}} / N$$

donde $a(i)$ es el promedio de las distancias del objeto al resto de las observaciones de su mismo clúster, y $b(i)$ es la distancia mínima a otro clúster distinto al de la observación i.

Los autores del artículo aplican el método de la *p-median* sobre un conjunto de datos recopilados por *Ross y Murphy (1999)*, en el que 38 sujetos clasifican 45 alimentos en tantas categorías como deseen, según la similitud percibida. Los datos se agregaron como proporciones de sujetos que no colocaron un par de alimentos en una categoría común (definiendo esto como disimilitudes en las que las proporciones representan los alimentos menos parecidos), conformando una matriz $45x45$ de distancias.

Las categorizaciones de un alimento pueden basarse en distintas situaciones, como por ejemplo un 'huevo', pudiendo clasificarse en desayuno, o producto lácteo, entre otras muchas.

Se obtuvieron óptimos globales para $2 \leq p \leq 14$ conglomerados. El número final de los p conglomerados se determinaron según el *Índice de Silueta (SI)*, que a pesar de que el promedio máximo corresponde a $p = 11$, los valores aumentan en mayor medida hasta $p = 8$, cuyo valor está comprendido en el rango establecido por *Kauffman y Reusseeuw (1990)*, 0.51 -0.70 (Tabla 1).

En la Tabla 2, se representan los ocho grupos en los que se ha realizado la agrupación. Los objetos en negrita son los 'centros' caracterizado por el problema, como la mediana en cada grupo, agregándose el resto de objetos a las medianas cercanas.

La elección del método de agrupamiento correcto a menudo causa una gran indecisión (métodos jerárquicos, métodos no jerárquicos, ...). En muchas aplicaciones prácticas, la decisión carece de un marco teórico inequívoco, por lo que se escoge aquella opción que mejor se ajuste al problema.

La agrupación *K-Means* y la agrupación clúster basada en modelos son coherentes si se desea analizar los datos siguiendo el agrupamiento por grupos (centroides). Sin embargo, la *p-median* es tomado por muchos investigadores como una técnica viable de agrupamiento. Esta consideración se debe a su robustez ante valores atípicos, aventajando a su vez al resto de métodos por su flexibilidad en la entrada de las matrices de distancia (simétricas o asimétricas, rectangulares o cuadradas, datos categóricos o de escala), frente a la mayoría de las técnicas de agrupamiento basada en modelos (matrices rectangulares de distancia).

Un hándicap del problem *p-median* es su gran esfuerzo computacional a medida que aumenta el número de objetos en el estudio. Como solución a esta problemática, la proposición de *Brusco y Köhn (2008b)* del método de tres etapas consigue óptimos globales para conjuntos con un tamaño $N = 1,400$ y máximo 30 grupos. La primera etapa, Sustitución de Vértices (VS) en general converge correctamente si el número de conglomerados (p) es menor o igual a 20.

Por lo tanto, si el estudio requiere de óptimos globales en un problema de agrupación en clústers, en función del tamaño de los datos, el método de la *p-median* sería óptimo.

A p-median problem with distance selection

El artículo 'A p-median problem with distance selection' de los autores *Stefano Benati* y *Sergio García* introducen una extensión del problema *p-median* en el clustering, a partir de la función de distancia entre unidades, calculada como la suma de distancias sobre las q variables más relevantes.

Para ello, introducen un problema de clustering donde nos dan un conjunto $U = \{u_I\}_{i=1}^{n}$ de unidades estadísticas, medidas mediante un conjunto de variables $F = \{f_k\}_{k=1}^{m}$. Ambos conjuntos son presentados en una matriz $V = [v_{ik}]$ donde v_{ik} es el valor de la característica f_k para la unidad u_i. El análisis clúster consta de dos pasos: el primero de ellos trata de establecer una medida de distancia d_{ij} entre cada par de observaciones (u_i, u_j), tras ello, se aplica un algoritmo de clustering con el que obtener una partición de los datos.

Uno de los métodos de clustering más importantes y sobre el que trata el presente artículo, el problema de la *p-median*, caracterizándose por la representación de cada clúster por la mediana de sus datos, minimizando la suma total de las distancias entre cada unidad y su mediana más cercana.

Con grandes conjuntos de datos, se obvia que no todas las variables contenidas son relevantes para el estudio, pudiendo distorsionar la función de distancia con el ruido generado por estas características. Por ello, se suelen considerar subconjuntos de la base de datos tal que $Q \subseteq F$, esto es que todas las variables del subconjunto están contenidas en F. Teniendo como objetivo la resolución de esta problemática de variables irrelevantes asociado al problema de la *p-median*, se propone un modelo de selección de Q características relevantes, un número de P medianas óptimas y cuyas distancias no solo depende de las medianas $P \subseteq U$, sino también de las características seleccionadas $Q \subseteq F$.

En esta sección se presenta la formulación del modelo anteriormente introducido sobre unos datos característicos. Se tiene un conjunto $U = \{u_I\}_{i=1}^{n}$ de unidades estadísticas, medidas mediante un conjunto de variables $F = \{f_k\}_{k=1}^{m}$ cualitativas u ordinales. Para datos cualitativos se representa por $[0-1]$, mientras que para datos ordinales con g ocurrencias o siguiendo una escala de Likert, g hace referencia a la dimensión de la escala.

Se supone que solo se considera relevante un conjunto de características $Q \subseteq F$, calculando la distancia solo con estas variables Q, por lo que la fórmula corresponde a

$$d_{ij} = \sum_{k=1}^{m} d_{ijk} z_k$$

definiendo z_k como variable binaria, $z_k = \begin{cases} 1, \text{si } z_k \text{ se encuentra en } f_k \in Q \\ 0, \text{en caso contrario} \end{cases}$

Una vez definida la función de distancia cuando se selecciona un subconjunto Q de variables, se definen dos variables de asignación necesarias para el modelo,

$y_j = \begin{cases} 1, \text{si la unidad } j \text{ es una mediana} \\ 0, \text{en caso contrario} \end{cases}$

$x_{ij} = \begin{cases} 1, \text{si la unidad } i \text{ se asigna a la mediana } j \\ 0, \text{en caso contrario} \end{cases}$

El modelo primario que se muestra es el siguiente:

$$\begin{aligned} & (F_0) = \min \sum_{i=1}^{n}\sum_{j=1}^{n}(\sum_{k=1}^{m} d_{ijk})x_{ij} & \\ \text{s.a.:} \quad & x_{ij} \leq y_j, & i,j = 1,...,n \\ & \sum_{j=1}^{n} x_{ij} = 1, & i = 1,...,n \\ & \sum_{j=1}^{n} y_j = p, & \\ & x_{ij} \geq 0, & i,j = 1,...,n \\ & y_j = \{0,1\} & j = 1,...,n \end{aligned}$$

este modelo ejecuta el algoritmo sobre todas las variables disponibles en los datos, ya que no se encuentra especificada la variable binaria (z_k), la cual hace referencia a las variables Q relevantes, por lo tanto, el modelo ampliado con esta suposición trataría:

$$
\begin{aligned}
& (F_1) = \min \sum_{i=1}^{n}\sum_{j=1}^{n}(\sum_{k=1}^{m} d_{ijk} z_k) x_{ij} \\
\text{s.a.:} \quad & x_{ij} \leq y_j, && i,j = 1,...,n \\
& \sum_{j=1}^{n} x_{ij} = 1, && i = 1,...,n \\
& \sum_{j=1}^{n} y_j = p, \\
& x_{ij} \geq 0, && i,j = 1,...,n \\
& y_j = \{0,1\} && j = 1,...,n \\
& z_k = \{0,1\} && k = 1,...,m
\end{aligned}
$$

Esta formulación de la función objetivo añade la variable z_k mediante la cual, solo se tienen en cuenta las distancias para el resultado, si esta variable es 1, es decir, si la variable seleccionada está en el subconjunto de características relevantes Q.

El problema no lineal, como el presente, lleva asignado problemas de computación debido a los términos cuadráticos. Por lo tanto, se intenta linealizar la función objetivo con la finalidad de computacionalmente, generar un modelo óptimo. Para ello se proponen en el artículo cuatro reformulaciones del modelo actual.

La primera de ellas hace referencia a la linealización directa de las variables cuadráticas añadiendo

$$w_{ijk} = x_{ij} z_k, \quad i,j = 1,...,n, \;\; k = 1,...,m$$

Esta nueva variable representa la asignación de {0,1} de la unidad i a la mediana j mediante k, obligando a la inclusión de tres nuevas premisas al modelo:

$$
\begin{aligned}
w_{ijk} \geq x_{ij} + z_k - 1, \quad & i,j = 1,...,n \;\; k = 1,...,m, \\
x_{ij} \in \{0,1\}, \quad & i,j = 1,...,n \\
z_k \in \{0,1\}, \quad & k = 1,...,m
\end{aligned}
$$

por lo que la formulación de esta propuesta de convertir en un modelo lineal el cuadrático sería:

$$(F_2) = \min \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} d_{ijk}w_{ijk}$$

s.a.:

$$x_{ij} \leq y_j, \qquad i,j = 1,...,n$$

$$\sum_{j=1}^{n} x_{ij} = 1, \qquad i = 1,...,n$$

$$\sum_{j=1}^{n} y_j = p,$$

$$\sum_{k=1}^{m} z_k = q,$$

$$w_{ijk} \geq x_{ij} + z_k - 1, \qquad i,j = 1,...,n \;\; k = 1,...,m,$$

$$x_{ij} \geq 0, \qquad i,j = 1,...,n$$

$$x_{ij} \in \{0,1\}, \qquad i,j = 1,...,n$$

$$y_j = \{0,1\} \qquad j = 1,...,n$$

$$z_k = \{0,1\} \qquad k = 1,...,m$$

El siguiente enfoque del problema sería la de una formulación de estilo *p-median*, para ello sólo se utilizarían las variables w_{ijk}, y_j, z_k.

$$(F_3) = \min \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} d_{ijk}w_{ijk}$$

s.a.:

$$\sum_{k=1} w_{ijk} \leq qy_j, \qquad i,j = 1,...,n, \tag{1}$$

$$\sum_{j=1}^{n} w_{ijk} = z_k, \qquad i = 1,...,n \;\; k = 1,...,m, \tag{2}$$

$$\sum_{\substack{t=1 \\ t\neq k}}^{m} w_{ijk} \geq (q-1)w_{ijk}, \qquad i,j = 1,...,n \;\; k = 1,...,m, \tag{3}$$

$$\sum_{j=1}^{n} y_j = p,$$

$$\sum_{k=1}^{m} z_k = q,$$

$$w_{ijk} \geq 0, \qquad i,j = 1,...,n \;\; k = 1,...,m,$$

$$y_j = \{0,1\} \qquad j = 1,...,n$$

$$z_k = \{0,1\} \qquad k = 1,...,m$$

estos cambios significativos, en la eliminación de algunas variables se explican de manera que, en la restricción (1) la observación i solo puede ser asignada a j si la variable es seleccionada como mediana. La restricción (2) en todo j, establece que la unidad i se asigna a alguna mediana, solo si la variable k se selecciona.

La siguiente reformulación se centra en el intento de reducir el tamaño del problema. Para ello se plante:

$$
\begin{aligned}
& (F_4) = \min \sum_{i=2}^{n} \sum_{j=1}^{i-1} h_{ij} \\
\text{s.a.:} \quad & x_{ij} \leq y_j, && i,j = 1,...,n, \\
& \sum_{j=1}^{n} x_{ijk} = 1, && i = 1,...,n \\
& \sum_{j=1}^{n} y_j = p, \\
& \sum_{k=1}^{m} z_k = q, \\
& w_{ijk} \geq 0, && i,j = 1,...,n \;\; k = 1,...,m, \\
& h_{ij} + M_{ij}(1 - x_{ij} - x_{ji}) \geq \sum_{k=1}^{m} d_{ijk} z_k, && 1 \leq j < i \leq n, \qquad (1) \\
& h_{ij} \geq 0, && i,j = 1,...,n \;\; j < i,
\end{aligned}
$$

Se introduce el concepto de h_{ij}, recoge el coste de asignar i a la variable j o viceversa, tomando valor 0 si no ocurre ninguna de las asignaciones.

Finalmente, se propone un modelo de formulación de radio aplicado al problema *p-median*. Se tiene una unidad u_i y un rasgo estadístico f_k expresado en la escala Likert con g grados. El procedimiento del modelo pasa a ser, en el primer paso, dado un cliente i, se ordenan crecientemente las distancias $\{d_{i1k}, d_{i2k}, ..., d_{ink}\}$ siendo D_{ik1} el menor coste, D_{ik2} el segundo menor coste y así sucesivamente.

Seguidamente se define,

$$
r_{ikt} = \begin{cases} 1, \text{si se selecciona la característica } k, \text{la unidad } i \text{ está asignada a una distancia} D_{ikt} \\ 0, \text{en caso contrario} \end{cases}
$$

El modelo resultante utiliza las variables definidas anteriormente y_j, x_{ij} y z_k:

$$
\begin{aligned}
& (F_5) = \min \sum_{i=2}^{n} \sum_{k=1}^{m} \sum_{t=2}^{G_{ik}} (D_{ikt} - D_{ik,t-1}) r_{ikt} \\
\text{s.a.:} \quad & x_{ij} \leq y_j, && i,j = 1,...,n, \\
& \sum_{j=1}^{n} x_{ij} = 1, && i = 1,...,n \\
& \sum_{j=1}^{n} y_j = p, \\
& \sum_{k=1}^{m} z_k = q, \\
& r_{ikt} + \sum_{j/d_{ijk} < D_{ikt}} x_{ij} \geq z_k, && i = 1,...,n, \;\; k = 1,...,m, \;\; t = 2,...,G_{ik}, \qquad (10) \\
& r_{ikt} \geq 0, && i = 1,...,n \;\; k = 1,...,m, \;\; t = 2,...,G_{ik}, \\
& x_{ij} \geq 0, && i,j = 1,...,n \\
& y_j = \{0,1\} && j = 1,...,n \\
& z_k = \{0,1\} && k = 1,...,m
\end{aligned}
$$

La función objetivo minimiza la distancia del cliente i si se selecciona la variable z. La nueva restricción característica del problema (10), obliga que, si se selecciona la característica k o el cliente i es asignado a una mediana j estando a una distancia $d_{ijk} \leq D_{ikt}$, o si no se cumple lo anterior, asigna a una distancia mínima D_{ikt} para $r_{ikt} = 1$.

Una vez se presentan las reformulaciones, F_5 (modelo de radio) tiene un menor número de restricciones y variables que los anteriormente descritos, por ello es un candidato óptimo en la resolución del problema. Para confirmar esta suposición, se prueban los modelos sobre datos generados con distintas dimensiones.

Primeramente, se definen los parámetros a utilizar, los valores que toman son:

- $n = 30, 50, 80$.
- $m = 8$(casos A) o $m = 12$(casos B).
- Los parámetros p y q tienen los siguientes valores:
  - Instancias A: $p = 2, 4, 6$, y $q = 2, 4, 6$.
  - Instancias B: $p = 2, 4, 6$, y $q = 3, 6, 8$.
- Tipo de datos: deterministas(H) o probabilísticos(P).
- Número de medianas: 2 o 4.
- Escala: binaria(L1) o Likert(L5).

El estudio computacional se ha propuesto sobre los datos, tal que para una escala dada (L1/L5), tipo de datos (A/B) y naturaleza de los datos (determinista/probabilística) hemos compactado la información que corresponde a las 27 instancias que tenemos cuando consideramos los diferentes números de unidades estadísticas ($n = 30, 50, 80$) y los nueves pares diferentes (p, q).

Esta combinación de datos se prueba con los modelos lineales propuestos. Se determina que el modelo lineal basado en radio, mejora al resto de los propuestos. Una vez esta conclusión, se determina que el modelo (F_4) es inestable en cuanto a su optimalidad, siendo en algunos problemas rápido para llegar al resultado mientras que en otros, su comportamiento es deficiente, considerandose el peor modelo frente al resto.

El éxito del modelo propuesto (5), referente a las variables de radio, alcanza el éxito entre el resto de los problemas debido a que su formulación se basa sobre unas pocas variables binarias. Así, se evalúa el tiempo del CPU para la resolución de los problemas, teniendo que el modelo (5) obtiene los mejores valores para cada una de las medidas, respecto al resto.

Finalmente, los autores han propuesto el modelo de la p-median con distancias entre las unidades estadísticas sobre un número q de variables originales F. Así, la linealización del problema principal cuadrático consigue una optimización considerable en el tiempo de computación que, sobre los cinco modelos diferentes destaca la formulación del problema de radio (F_5), siendo un importante punto de partida sobre estudios o agrupaciones del ámbito sociológico y de encuestas, en los cuales se recogen muchas variables sin esclarecer la importancia de cada una. Se trata de un buen modelo para el estudio de variables binarias u ordinales por el uso que hace el modelo de la distancia.

Los autores suscitados por una aplicación de agrupación sobre encuestas, proponen este modelo *p-median* con distancias entre las unidades estadísticas sobre un número q de variables originales sobre el total de variables m. El hecho de la gran cantidad de variables recogidas en las encuestas sociológicas obliga al investigador seleccionar un número menor de variables suponiéndose relevantes para el estudio, por ello, el modelo propuesto que selecciona q variables de un total de F, además del buen manejo de variables binarias u ordinales por el uso que hace el modelo de la distancia, mayoritariamente la tipología de variables de encuestas, se considera una perfecta herramienta para el fin perseguido por dichos autores. Así, la linealización del problema principal cuadrático consigue una optimización considerable en el tiempo de computación que, sobre los cinco modelos diferentes destaca la formulación del problema de radio (F_5),

siendo un importante punto de partida sobre estudios o agrupaciones del ámbito sociológico y de encuestas, en los cuales se recogen muchas variables sin esclarecer la importancia de cada una.

Según lo expuesto en el artículo anterior acerca de la formulación de radio, el cuál consigue eliminar variables redundantes de la función objetivo, sustituyendo las variables de asignación de elementos a clúster, por variables radiales. Las distancias calculadas para resolver el modelo de la *p-median* no solo se basa en la comprendida entre los objetos, si no, también la distancia entre las variables elegidas sobre el total.

Para ello, se propone como tema de aplicación la selección de variables relevantes en análisis de imágenes. Este problema es singular, porque cada píxel de una imagen puede corresponder a una variable, por lo que a la hora de formar grupos, la no inclusión de variables irrelevante en el modelo, es de vital importancia.

Mixed integer linear programming and Heuristic methods for feature selection in clustering

El artículo publicado en 2017 por los autores *Sergio García*, *Stefano Benati* y *Justo Puerto*, "Mixed integer linear programming and Heuristic methods for feature selection in clustering", estudia un problema recurrente en el análisis clúster como es la selección óptima de variables y evitar incluir en el estudio variables irrelevantes que pueden alterar el resultado. Esta acción se propone a partir de métodos heurísticos.

El objetivo del análisis clúster es el de agrupar objetos similares en un clúster, buscando que los objetos agrupados sean muy homogéneos entre sí, agrupando a su vez los objetos disímiles en distintos clúster, consiguiendo grupos heterogéneos entre sí.

La problemática radica cuando la heterogeneidad entre los diferentes clúster no es tal, dando lugar a clúster inconsistentes, siendo una de las principales causas la inclusión de variables poco relevantes en el proceso.

La literatura sobre este problema se ha centrado en tres métodos para seleccionar o rechazar. Múltiples autores han sugerido el uso de índices de 'clusterabilidad'. Entre los que se encuenta el índice TOPRI (Total Pairwise Rand Index)[19], el cual selecciona las variables más influyentes de parada. Otro índice de 'clusterabilidad' trata de añadir una restricción la cual equilibra la correlación con la reducción de la variabilidad, también se expone como índice de 'clusterabilidad' un procedimiento de normalización y selección de variables mediante optimización.

Un segundo enfoque es dirigido a la optimización de la función de verosimilitud, suponiendo la distribución de los datos como multivariantes. Para ello, se necesita calcular todas las medias, varianzas y covarianzas. La optimización de la función de verosimilitud debe realizarse varias veces en el caso de selección de características, por lo que el tiempo de cálculo aumente en gran medida. En ciertos estudios, se ha añadido un término de penalización a la función de máxima verosimilitud con el que se consigue rechazar ciertas variables.

Mientras que el tercer ámbito de estudio del problema nace en la conservación de la estructura del problema de selección de variables, sustituyendo la función de máxima verosimilitud por otra más sencilla. Para ello, se ha estudiado la inclusión de la función objetivo del modelo *k-Means* añadiendo ciertas penalizaciones, teniendo de esta manera resultados óptimos.

El estudio se articula en dos bloques, en el primero se aborda el problema de la selección de variables en el análisis clúster como un problema de programación entera, mientras que en el segundo bloque se resuelve esta problemática a partir de dos algoritmos heurísticos.

A continuación, se hace un repaso acerca del problema de clustering, teniendo $U = \{1, ..., n\}$ objetos, cuyas variables o características $V = \{1, ..., m\}$ siendo alguna de estas relevantes y otras en las que sus valores no son significativos para la pertenencia a un grupo y los centroide seleccionados en un análisis preliminar $R = \{1, ..., r\}$. Teniendo $i \in U$, $k \in R$, $j \in V$, siendo d_{ij} la distancia entre i y k medida por la variable j. La asignación de los objetos U a uno de los centros R se determina por la distancia más corta, siendo $D(Q) = \sum_i d_{i,k(i)}(Q)$ la suma de todas las distancias entre los objetos y los centros, seleccionando así las variables cuando se minimiza $D(Q)$, añadiendo la restricción $|Q| = q$ (siendo q un parámetro fijado con anterioridad). Los investigadores denominan a este nuevo problema como, *'q-variable selection'*.

Para una Q fija, se resuelve mediante el problema de la *p-median*, siendo el problema *q-variable selection* una simplificación de este, ya que R son fijos.

El índice $D(Q)$ está relacionado con la variabilidad dentro de los clúster, si se estandarizan las variables con la puntuación z, la función objetivo es equivalente a la minimización de la variabilidad dentro de los grupos. Siendo s_{ij} el valor de j para la unidad i, $\mu = \frac{1}{n}\sum_{i=1}^{n} s_{ij}$ sea la media j, teniendo que la variabilidad aportada por j es:

$$TSS_j = \sum_{i=1}^{n}(s_{ij} - \mu_j)^2$$

El caso en el que las unidades se dividen G_k, $k = \{1, ..., r\}$, y la media de la característica j de cada centro de conglomerado es $r_{kj} = \frac{1}{G_k}\sum_{i \in G_k} s_{ij}$, para $j = 1, .., m$; $k = 1, ..., r$, por lo tanto, la variabilidad dentro de los clúster se explica como

$$WSS_j = \sum_{i=1}^{n}(s_{ij} - r_{k(i),j})^2$$

mientras que la variabilidad entre clúster es definida:

$$CSS_j = \sum_{i=1}^{n}(r_{k(i),j} - \mu_j)^2$$

Una vez explicada la casuística de la selección de variables y los índices utilizados para ello, así como la minimización de la variabilidad entre y dentro de los clúster, los autores describen el problema de programación lineal con el que llevar a cabo la selección de las q variables más relevantes para el estudio.

A continuación, los autores plantean dos formulaciones distintas para el problema de selección de características Q.

La primera de ellas añade variables de decisión para la selección de variables como para la asignación de objetos a los grupos, generándose un número cuadrático de variables binarias para las asignaciones.

Mientras que, para la segunda formulación de programación lineal entera del problema, la propuesta es la sustitución de las variables binarias de asignación por variables de radio, reduciendo la cantidad de variables del problema, simplificando el cálculo computacional del modelo. Donde para el primer modelo se tiene:

- $z_j, \quad j = 1, ..., m$; representa si la característica j se elige o no, es decir, $z_j = 1$ si $j \in Q$, $z_j = 0$, en caso contrario.
- $x_{ik}, \quad i \in U, k \in R$, son las variables de asignación global de la unidad i al centro del clúster k, es decir, $x_{ik} = 1$, si la unidad i es asignada al centro de clúster k, $x_{ik} = 0$, en caso contrario.
- $w_{ijk}, \quad i \in U, j \in V, k \in R$; son las variables locales de asignación de la unidad i al centro del clúster k utilizando la característica j, siendo $w_{ijk} = 1$ si la unidad i es asignada al centro de clúster k y la característica j es elegida, $w_{ijk} = 0$, en caso contrario.

$$(P_1) : f(z, x, w) = \min \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{r} d_{ijk} w_{ijk} \tag{1}$$

$$\text{s.a.:} \quad \sum_{j=1}^{m} w_{ijk} = q x_{ik}, \tag{2}$$

$$\sum_{k=1}^{r} x_{ik} = 1, \tag{3}$$

$$\sum_{k=1}^{r} w_{ijk} \leq z_j, \tag{4}$$

$$\sum_{k=1}^{m} z_j = q, \tag{5}$$

$$w_{ijk} \in \{0, 1\}$$

$$x_{ij} \in \{0, 1\}$$

$$z_j \in \{0, 1\}$$

La formulación del problema establece que en la restricción (3), cada unidad i es asignado a un clúster k. Una asignación local (i, k) solo es factible si la variable j ha sido seleccionada expresado en la restricción (4). Mientras, la restricción (5) fija el número de variables en q.

En el caso de no tener varianzas iguales las variables, los investigadores establecen un doble objetivo, por una parte, minimizar $\sum_{k \in Q} WSS_k$ y maximizar $\sum_{k \in Q} CSS_k$. Este biobjetivo puede simplificarse añadiendo un límite tal que:

$$\sum_{k=1}^{m} CSS_{kzk} \geq K \tag{10}$$

Siendo K elegido por el investigador con anterioridad.

Así, para tratar de evitar variables correlacionadas en el estudio, se propone la inclusión de la siguiente restricción,

$$|p_{ij}| \leq (\min\{WSS_i, WSS_j\})^2 \tag{11}$$

Siendo p_{ij} la correlación entre las variables i y j, si no se cumple la desigualdad solo puede seleccionarse la variable i o j.

En el caso de necesitar que el tamaño de los clústers sea igual, se propone una restricción en el que se permite un rango de cardinalidad entre l y u,

$$l \leq \sum_{i \in U} x_{ij} \leq u \tag{13}$$

Debido a la flexibilidad del modelo, a su vez se puede eliminar los valores atípicos entre los datos de estudio utilizando el método *k-Means* recortado, variando del modelo original de *k-Means*, eliminando un porcentaje α de las unidades estadísticas más alejadas, descartándose del cálculo de las medias,

$$\sum_{i \in U} \sum_{j \in R} x_{ij} = \lfloor n(1 - \alpha) \rfloor \tag{14}$$

A continuación, se expone el problema que sustituye las variables binarias de asignación por las variables de radio, consumiendo menos tiempo computacional.

Este problema elimina de su análisis las variables w_{ijk} por h_{ijt}, donde,

$$h_{ijt} = \begin{cases} 1, & \text{si se selecciona la variable } j \text{ y si la unidad } i \text{ se asigna a un centro } k \text{ tal que } d_{ijt} \geq D_{ijt} \\ 0, & \text{en caso contrario} \end{cases}$$

Considerando la Figura 1, se define que $d_{iju} = d_{ijv} = D_{ij1}$, porque u y v están en la misma circunferencia

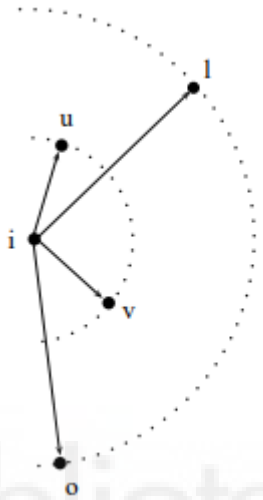

Figure 1: Descripción del radio de puntos equidistantes

Para un par $i \in U, j \in V$, el término correspondiente de la función objetivo debe reescribirse:

$$P_2 : \min \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{t=1}^{g(i,j)} (D_{ijt} - D_{ij,t-1}) h_{ijt} \tag{1}$$

$$\text{s.a.:} \quad \sum_{k=1}^{r} x_{ik} = 1, \qquad \forall i, \tag{2}$$

$$\sum_{j=1}^{m} z_j = q, \tag{3}$$

$$h_{ijt} + \sum_{\{k | d_{ijk} < D_{ijt}\}}^{n} x_{ik} \geq z_j, \qquad \forall i, \forall j, \forall t \geq 1, \tag{4}$$

$$h_{ijt} \leq 0, \qquad \forall i, \forall j, \forall t, \tag{5}$$

$$x_{ik} \geq 0 \qquad \forall i, \forall j, \tag{6}$$

$$z_j \in \{0, 1\} \qquad \forall k. \tag{7}$$

La formulación del problema establece que, si se selecciona la característica j ($z_j = 1$) y la unidad i no está asignada al clúster k, h_{ijt} toma valor 1. Por otra parte, si $z_j = 0$, para alguna característica j, entonces las variables de radio $h_{ijt} = 0$ para todo i, ya que la función objetivo busca la minimización del problema, mientras que si $z_j = 1$, para cualquier i existe una solución óptima para x_{ik}.

La resolución de los algoritmos de selección de variables conlleva un tiempo de cálculo creciente exponencialmente debido a que las variables óptimas z,y las asignaciones óptimas de las variables x se realizan con

cálculos simultáneos, pero si la resolución es realizada de manera independiente, es decir, de forma separada, los problemas son resolubles.

Los autores han reformulado el problema para conseguir una solución factible de asignación, siendo $k(i)$ el centro del clúster k donde $x^*_{i,k(i)}$.

$$\min \sum_{i=1}^{n} (\sum_{j=1}^{m} d_{i,j,k(i)} w_{i,j,k(i)}) \tag{22}$$

$$\text{s.a.:} \quad \sum_{j=1}^{m} w_{i,j,k(i)} = q, \qquad \forall i / x_{i,k(i)} = 1, \tag{23}$$

$$w_{i,j,k(i)} \leq z_j, \qquad \forall i, \forall j \tag{24}$$

$$\sum_{j=1}^{m} z_j = q, \tag{25}$$

$$0 \leq w_{i,j,k(i)} \leq 1, \qquad \forall i, \forall j \tag{26}$$

$$0 \leq z_j \leq 1 \qquad \forall j. \tag{27}$$

Debido a las restricciones (23) y (25), se puede reformular la restricción (24) como una igualdad, tal que:

$$j : q = \sum_{j=1}^{m} w_{i,j,k(i)} \leq \sum_{j=1}^{m} z_j.$$

Por lo que, según lo propuesto por los autores, para $z_{j(t)} = 1$, $t = 1, ..., q$ como solución óptima. Finalmente, para los valores de z y x existe una solución w con valores enteros.

Si el vector de las variables z es fijo, las distancias entre los objetos y los clúster son calculadas de manera sencilla asignando óptimamente la unidad i al centro del conglomerado k.

En función del problema propuesto, un proceso heurístico puede alternar entre dos subrutinas, teniendo:

La subrutina *Best-Assignment* puede formularse:

- Paso 1: Para todo $i \in U, k \in R$ sea $c_{ik} = \sum_{j \in Q} d_{ijk}$.
- Paso 2: Para todo $i \in U$, que x_{iw} si $c_{iw} = \min\{c_{ik} | 1 \leq k \leq r|\}, x_{iw} = 1$ en caso contrario.
- Paso 3: Sea $D(Q) = \sum_{i \in U} \sum_{k \in R} c_{ik} x_{ik}$.

Subrutina *Best-Assignment*, empieza con alguna z fija hasta encontrar la asignación óptima. Así, calcular la z óptima para esa x y repetir hasta no conseguir una mejora.

Mientras, la subrutina *Best-Variable*:

- Paso 1: Para todo $j \in V$, sea $b_j = \sum_{i,k} d_{ijk} x_{ik}$.
- Paso 2: Ranking b_j en orden creciente: $b_{j(1)} \leq ... \leq b_{j(m)}$.
- Paso 3: Sea $j(i) \in Q$, si y solo si, $i \leq q$.
- Paso 4: Sea $C_X(Q) = \sum_{i=1}^{q} b_{j(i)}$.

Por otra parte, la subrutina *Best-Variables* calcula un conjunto óptimo de variables Q para una asignación dada, a los clúster X.

El problema puede ser descompuesto en las dos subrutinas anteriormente descritas, para lograr un problema polinomialmente resoluble. En un primer paso con la subrutina *Best-Variables* desde un conjunto de variables Q^0 se calcula la mejor asignación de X^0, calculando tras la asignación de X^0 las variables óptimas Q^1. En el caso de $Q^1 \neq Q^0$, se calcula una nueva asignación de X^1 hasta tener $Q^t = Q^{(t-1)}$.

Tras la exposición de las subrutinas de selección óptima de variables y de asignación, en el artículo se presenta el modelo *q-vars*, el cual tiene semejanza con el algoritmo *k-Means*.

A grandes rasgos, el modelo propuesto *q-vars*, comienza con una selección aleatoria de variables desde las cuales, se calcula alternativamente las asignaciones óptimas de objetos y variables. Analíticamente:

- Paso 1: Inicio aleatorio: Selección aleatoria de un conjunto de variables Q^0 con $t := 0$
- El proceso se repite hasta encontrar un óptimo local.
  - Paso 2: (Asignación de unidades) Para un Q^t dado, llame a la subrutina *Best-Assignments* para calcular el X^t óptimo y el $C_{Q^t}(X^t)$.
  - Paso 3: (Selección de variables) Para un X^t dado, llame a la subrutina *Best-Variables* para calcular el $Q^t + 1$ y $C_{X^t}(Q^{t+1})$ óptimos y actualice $t := t + 1$
- Paso 4: $C^{best} = \min\{C^{best}, C_{Q^t}(X^t)\}$, actualizar Q^{best} en consecuencia.
- Paso 5: $s = s + 1$. Si $s \leq s^{max}$ se vuelve al paso 1.

El siguiente método definido se llama *Add-and-Drop*[20], ya que la solución óptima la consigue añadiendo y eliminando variables. Comienza con un conjunto de soluciones $Q \subseteq V$, si la función objetivo disminuye, se añade una nueva variable de $V - Q$ a Q y se elimina una variable de Q.

Una vez desarrollados los dos problemas MILP (P1, P2) y los dos métodos heurísticos (*q-vars*, *Add-and-Drop*), los autores evalúan la complejidad de resolución computacional de estos 4 modelos en la selección de variables significantes y la asignación óptima.

La primera prueba comparativa que se realiza en el estudio se define, $n = 15$ unidades estadísticas divididas en 4 grupos, cada uno de los grupos está descrito por 20 variables relevantes, para $m = 50, 100, 500, 1000$ (variables binarias del modelo) se añaden $m - 20$ variables ruidosas.

Computacionalmente se requiere que los problemas MILP dispongan para la resolución 7200 segundos, mientras que, para los modelos heurísticos, el criterio restrictivo es $s^{max} = 100$.

Los resultados se recogen en la *Tabla 3*, donde se concluye que el algoritmo con mejor rendimiento es el heurístico *q-var*, para el que, fijando $q = 20$ coincidiendo con el valor real, o cuando erróneamente $q = 10$, $q = 40$ encuentra la solución óptima. Por su parte, el algoritmo *Add-and-Drop* es mucho más lento y solo encuentra el óptimo cuando $q = 20$. Los modelos LP llevan asociado un tiempo de resolución muy bajo, excepto para $q = 40$, $m \geq 500$. Para la reducción de dimensionalidad se ejecuta otra prueba, bajo la premisa de que todas las variables son significativas, pero variando lo discriminante de cada una.

Las unidades estadísticas pertenecen a dos grupos G_k,para $j = 1, \ldots, m$. Los resultados obtenidos en la *Tabla 4*, en este caso, de los métodos heurísticos por tiempo y precisión, *q-vars* mejora al modelo *Add-and-Drop*. Así, el modelo P_2 con variables de radio es mejor en estos casos que P_1. Establecen una tercera prueba, en la que los grupos no están perfectamente separados, si no, que se superponen entre ellos. Ante esta situación, la *Tabla 5* muestra como los algoritmos MILP no consiguen encontrar un valor óptimo dentro de las dos horas establecidas como límite, mientras que con los modelos heurísticos aumenta el tiempo y el número de iteraciones considerablemente. Por lo que, en aplicaciones sobre datos reales, donde se superpongan los grupos no podrán utilizarse los modelos MILP, utilizando, tenor a los resultados la heurística *q-vars*, ya que, como se ha visto en los resultados anteriores requiere de menos iteraciones para encontrar la solución óptima, es rápido y resuelve problemas a los que los modelos MILP no pueden. El uso de la técnica de

clustering es uno de los principales instrumentos utilizados con el fin de seleccionar variables significativas en posteriores análisis. Para esta consecución se determinan tres pasos imprescindibles, en primer lugar, la estandarización de las variables y evitar posibles distorsiones entre las distintas escalas. Seguidamente se resuelve el problema *q-vars* (selection variables) con el fin de seleccionar un conjunto óptimo de Q variables. Una vez este paso, aplicar el algoritmo de clustering sobre las Q variables óptimas.

Los autores proponen dos enfoques para llevar a cabo la estandarización de las variables, en un primer caso, con la puntuación z, restando la media de la distribución al valor de la variable entre su varianza, mientras que el segundo enfoque propuesto añade un factor de corrección sobre el *z-score*, intentando preservar el índice de agrupabilidad anteriormente descrito. Referente a la selección óptima de variables en el *Paso 2*, siguiendo la subrutina propuesta en el método *q-vars*, la subrutina *Best-Variables*, se ejecuta sobre distintos valores de q antes de seleccionar el mejor. La aplicación de clustering sobre las Q variables óptimas (*Paso 3*) se fundamenta en dos algoritmos populares como *k-Means* y de *maximización de expectativas* (*EM*).

Para tratar de medir el rendimiento de los métodos descritos, se proponen tres medidas.

Recuperación de clusters, mediante el índice ARI siendo 1 cuando la recuperación es perfecta y se acerca a 0 cuando la recuperación es igual a la elección aleatoria de clusters. [1]

La precisión es medida por el número de variables significativas seleccionadas entre la cardinalidad del subconjunto seleccionado. Un valor de 1 significa que todas las variables son relevantes.

Por último, la recuperación es medida como el número de variables relevantes del subconjunto seleccionado entre el total de las variables relevantes. Una vez expuesto los distintos algoritmos de selección y agrupación, así como las medidas de rendimiento, se realiza una pequeña prueba sobre datos simulados.

Las comparaciones entre los distintos métodos explicados en este artículo se realizan sobre una muestra de datos simulada, la cual contiene cuatro clúster de igual tamaño con 62 unidades cada uno. Se añaden variables de enmascaramiento a las relevantes, cumpliendo las de enmascaramiento que son normales independientes. Otro caso propuesto es que las variables de enmascaramiento son normales con media 0 y matriz de covarianzas cuyos términos en la diagonal son iguales a 1 y off-diagonales igual a 0.5. En un tercer escenario, las variables siguen distribuciones uniformes {0, 1}. La última problemática propuesta, es que las variables tienen distribuciones gamma con parámetros de localización y escala iguales a 1. Los resultados recogidos en la *Tabla 6*, muestran que, respecto a la capacidad de agrupación de los distintos algoritmos, los mejores ARI lo tienen los algoritmos *q-vars* (q-v1 que utiliza EM, y el algoritmo q-v2 que utiliza k-means). En la selección de variables, los algoritmos que seleccionan variables relevantes con alta probabilidad se asocian a una baja recuperación (probabilidad que alguna variable relevante se descarte), siendo *sb-red*, propuesto en [2], el algoritmo con mayor precisión.

En la *Tabla 7* y *8*, recogen los resultados para las estructuras de las variables de enmascaramiento. Cuando existe normalidad en los datos, el ARI con mayor puntuación lo tiene el algoritmo *cvs*, mientras que, para datos no normales, los métodos basados en k-means son menos sensibles.

Las *Tablas 9* y *10* condicionadas al grado de separación entre clusters, cuando este es bajo, el algoritmo más discriminante es *qv-1* basado en *k-Means*. Por último, las *Tablas 11* y *12* presentan datos condicionados a la relación entre variables de enmascaramiento y verdaderas. Para un ratio de 1, el mejor método corresponde a *qv-2*, mientras que para un ratio de 3, el algoritmo con mejor coeficiente es *qv-1*, debido a que datos con ruido afectan al algoritmo EM, al contrario de *k-Means*.

Según lo expuesto en el artículo, no existe un algoritmo superior al resto, ya que esto variará en función de la calidad de los datos (solapamiento del clúster, datos no normales, variables de enmascaramiento...), proponiendo a la combinación q-vars/k-Means como herramienta ante datos complicados.

Solving Capacitated P-median Problem by Hybrid K-Means clustering and FNS algorithm

En el siguiente artículo, los autores *Payman Kaveh, Ali Sabzevari Zadeh y Rashed Sahraeian* tratan de resolver el problema de la *p-median capacitada (CPMP)* a través de un algoritmo híbrido mediante *k-Means*

y el algoritmo de *búsqueda de vecindario fijo (FNS)*. El problema de la *p-median capacitada* consiste en la apertura de instalaciones en emplazamientos óptimos, minimizando la distancia entre los usuarios y la nueva planta. La diferenciación entre *CPMP* y el problema de la *p-median* radica en que el primero añade una restricción de capacidad para cada instalación.

Al tratarse de un problema de dificultad NP, el coste computacional para su resolución puede ser inasumible, por lo que se considera el uso de técnicas heurísticas para aproximar a la solución óptima.

Una de las mejores técnicas heurísticas para resolver el problema p-median capacitado es el algoritmo de *búsqueda por vecindario variable (VNS)*. El algoritmo de *búsqueda de vecindario fijo (FNS)* no es capaz de obtener la solución inicial, pudiendo interferir en el resultado final, por lo que se propone la solución inicial a través del algoritmo de agrupación *k-Means*. Tras esta primera acción, en iteraciones sucesivas, *FNS* las soluciones obtenidas se mejor a partir del intercambio de las ubicaciones de las instalaciones.

Matemáticamente, se describe el modelo teniendo que $N = \{1, \ldots, n\}$ número de clientes y, $J = \{1, \ldots, m\}$ emplazamientos posibles para la ubicación de instalaciones. Mientras c_{ij}, d_i,b_j representan la distancia entre el cliente i-ésimo ($i \in N$) y la ubicación j-ésimo ($j \in J$), la demanda del cliente i-ésimo y la capacidad de la instalación ubicada en j. Las variables de decisión que se introducen en el problema son

$$y_j = \begin{cases} 1, \text{ si una instalación está ubicada en } j, \forall j \in J \\ 0, \text{ en caso contrario} \end{cases}$$

$$x_{ij} = \begin{cases} 1, \text{ si el cliente } i, \forall i \in N, \text{ es asignado a la localización } j, \forall j \in J \\ 0, \text{ en caso contrario} \end{cases}$$

El modelo matemático del *CPMP* es:

$$\min \sum_{i \in N} (\sum_{j \in J} c_{ij} x_{ij}) \tag{1}$$

$$\text{s.a.:} \quad \sum_{j=J}^{n} x_{ij} = 1, \qquad \forall i \in N, \tag{2}$$

$$\sum_{j \in J} y_j = p, \tag{3}$$

$$\sum_{i \in N} d_i x_{ij} \leq b_j y_j, \qquad \forall j \in J, \tag{4}$$

$$x_{ij} \in \{0,1\}, y_j \in \{0,1\}, \quad \forall i \in N, \forall j \in J \tag{5}$$

La función objetivo minimiza la distancia total entre las instalaciones y los puntos de demanda. La restricción (2) obliga a que todo cliente sea asignado a una instalación, mientras que la (3) restringe la apertura de nuevas instalaciones hasta un total de p. Finalmente, la restricción (4) garantiza que la capacidad de cada instalación no tenga más demanda de la que puede cubrir.

Una vez explicado el problema de la p-mediana con capacidad, se aborda el método híbrido heurístico para resolver CPMP en un menor tiempo, siendo este una mezcla de *k-Means* y *FNS*.

Uno de los algoritmos más ampliamente utilizados en el aprendizaje no supervisado es el algoritmo *k-Means*, clasificando los objetos a un determinado número de clúster fijados a priori, definiendo cada clúster sobre su centroide. De manera resumida, en un primer paso se seleccionan k centroides al azar como centros iniciales de los clúster, seguidamente se asignan los clientes a los centroides más cercanos, creando k nuevos clúster. Se recalculan los clúster creados y sus centroides, finalizando el proceso una vez no se reasignen los objetos a clústers.

De forma análoga los valores de k y p en *CPMP* son iguales. Debido a la sensibilidad del algoritmo a los centroides seleccionados inicialmente, se considera necesario ejecutar varias veces el algoritmo con distintos centroides, para ello se propone un criterio con el que se consigue seleccionar el mejor.

$$Min\ Z = \sum_{i=1}^{n}\sum_{j=1}^{k} ||X_i^{(j)} - C_j||^2$$

$X_i^{(j)}$ muestra la ubicación del cliente i asignado por el centroide C_j al clúster j, minimizando la distancia entre el cliente y el centroide de pertenencia. El algoritmo híbrido propuesto utiliza la agrupación $k-Means$ como paso inicial en el algoritmo de *búsqueda de vecindario fijo (FNS)*. A continuación, se explica el algoritmo de *búsqueda de vecindad fijo (FNS)*, la casuística seguida por el algoritmo es, en primer lugar, se identifican los k' valores (contiene las soluciones que difieren de la ubicación actual), tras ello, mediante el intercambio de la k-ésima vecindad de la solución inicial S, evaluar en el modelo y comprobar si el valor de la función mejora el inicial, finalizando el algoritmo cuando no se produzca una mejora. El algoritmo FNS modificado difiere del algoritmo FNS puro en tres grandes rasgos. El primero de ellos corresponde a la posible inclusión de candidatos inadecuados, para ello se eliminan las k' instalaciones de la solución actual, especificando para las $p - k'$ restantes las h ubicaciones candidatas más cercanas. En segundo lugar, establecen un nuevo criterio de parada, para ello, si para cada k' eliminada de la solución actual, la mejor solución no mejor se activa el criterio de parada y finaliza el algoritmo. Terceramente, el algoritmo modificado FNS utiliza una lista *Tabu (LS)*, para lograr la solución óptima en un tiempo considerable. Son propuestos dos métodos para omitir emplazamientos ineficientes, disminuyendo el tiempo de resolución y la solución óptima es encontrada con mayor probabilidad. El primero de los métodos se realiza eliminando k' ubicaciones aleatorias de la solución actual, para las $p - k'$ restantes, se especifican las h ubicaciones más cercanas. En el paso siguiente, se evalúan las h posibles ubicaciones. El segundo método, se fundamenta en la idea de que la mayoría de los clientes no se encuentran en la frontera de una región delimitada, si no en la zona interior. Para ello, se omiten algunos candidatos, teniendo:

$$(X_{min}, X_{max}) = (Min(X_j), Max(X_j)) \quad \forall j \in J \tag{7}$$

$$(Y_{min}, Y_{max}) = (Min(Y_j), Max(Y_j)) \quad \forall j \in J \tag{8}$$

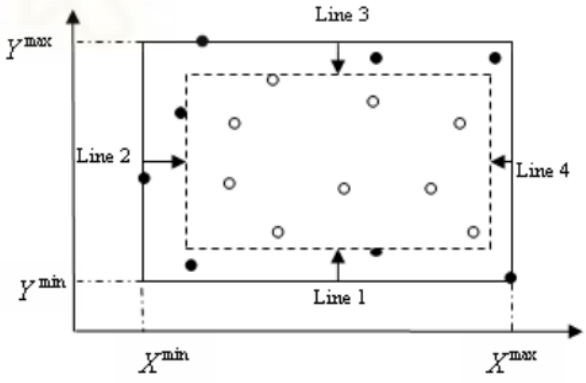

Figure 2: Ejemplo de omisión de sitios candidatos

Por lo tanto, se traza el área rectangular, donce para cada línea i, $i = 1, 2, 3, 4$, N_i, $i = 1, 2, 3, 4$, se especifican los sitios candidatos más cercanos a la línea i. Acto seguido, estos sitios se omiten.

Una vez introducidos los dos métodos para omitir del análisis ubicaciones ineficientes, los autores proponen dos criterios de finalización del algoritmo. El primer caso, corresponde a un número máximo de iteraciones, mientras que el segundo criterio, evalúa cada k' instalaciones diferentes eliminadas no mejora la solución, el algoritmo finaliza, optimizado este paso ya que no reevalúa las k' instalaciones debido a la memoria Tabu.

Con el fin de comprobar la eficiencia del algoritmo propuesto, se ha probado en dos conjuntos de instancias, el primero contiene 10 instancias de problemas de tamaño $n = 50$ y $p = 5$ y un segundo conjunto de 10

instancias de problemas de tamaño $n = 100$ y $p = 10$. Se crea un indicador para medir el porcentaje de desviación de la mejor solución conocida como:

$$dev = \frac{\text{Solución del algoritmo} - \text{Solución óptima}}{\text{Solución óptima}} \cdot 100$$

Se compara el algoritmo propuesto entre otros algoritmos como *RR*, *SS* y *PR+SS* y *VNS*, observándose que de los $P_1, \ldots, P_{20}$, la desviación de los resultados respecto a la solución conocida es de 0. En la mayor parte de estos, los algoritmos *RR*, *SS* y *PR+SS* tienen un rendimiento menor que el algoritmo propuesto, observando que en 10 ejecuciones el algoritmo propuesto en el artículo consigue encontrar la solución.

Finalmente, se destaca que gracias a la solución inicial aportada por el algoritmo de *k-Means* para el algoritmo *FNS* modificado, y a que dicho algoritmo omite candidatos inadecuados, y evita la reevaluación de candidatos ya que utiliza lista tabú, los resultados son óptimos.

En el presente artículo se ha introducido el algoritmo de agrupación *k-Means* como paso inicial para el cálculo de una solución inicial. Entendiendo el problema *p-median* como método de agrupación, se encuentran diversas similitudes entre ambos métodos. El problema *p-median* selecciona un número de p medianas como centro de cada grupo, por su parte, el algoritmo *k-Means* establece un número de k clúster con k centroides como centro de cada uno.

Otra de las similitudes es el uso de una medida de distancia para la agrupación de objetos. El algoritmo *k-Means* asigna los N objetos a cada grupo, minimizando la distancia de estos y los centroides que caracterizan cada grupo, análogamente, el problema *p-median* asigna los $N - p$ objetos restantes a cada mediana representativa de cada grupo minimizando la distancia entre estos.

The p-median problem: A survey of metaheuristic approaches

Los autores *Pierre Hansen, Jack Brimberg, Nenad Mladenovi'c y José A. Moreno Pérez* ofrecen una serie de técnicas metaheurísticas para la resolución de problemas de localización de instalaciones discretas, siendo en este caso, el problema de la *p-median (PMP).*

Los problemas de localización tratan de proponer nuevas posiciones de nuevas instalaciones, relacionadas con la distancia entre las ya existentes. El tamaño del problema de localización, a menudo, ha sido causa de problemas relacionado con el tiempo computacional para su resolución. Numerosos casos han resultado ser demasiado grandes para encontrar la solución. *Kariv y Hakimi (1969)* demostraron que el problema *p-median* es NP-Duro, por lo que la búsqueda de la resolución del problema mediante técnicas heurísticas que proporcionan una solución óptima o aproximada, que sin demostrar su optimalidad es una herramienta aceptada. Siendo en la última década, donde más se ha investigado sobre las técnicas metaheurísticas para la resolución de problemas NP-Duro, por ello es habitual recurrir a estas técnicas heurísticas o metaheurísticas.

A continuación, se muestra matemáticamente el problema *p-median (PMP)* en el ámbito de la localización de instalaciones. Este problema se utiliza en una gran cantidad de campos, como el análisis de conglomerados, la localización de almacenes, etc. En primer lugar, la definición de los conjuntos que se evalúan en el modelo, teniendo un conjunto L de m instalaciones, el conjunto U de k usuarios o puntos de demanda, y una matriz $D = nxm$ con las distancias recorridas por el usuario situado en i a la instalación situada en j, para todo $j \in L$, $i \in U$, buscando minimizar los costes o distancias. Añadiendo dos conjuntos de variables de decisión como:

$$y_j = \begin{cases} 1, \text{ si se abre una instalación en} j, \forall j \in L \\ 0, \text{ en caso contrario} \end{cases}$$

$$x_{ij} = \begin{cases} 1, \text{ si el cliente } i \text{ es atendido por la instalación } j, \forall i \in U, \forall j \in L \\ 0, \text{ en caso contrario} \end{cases}$$

Resultando el modelo de programación entera,

$$\min \sum_i (\sum_j d_{ij} x_{ij}) \tag{5}$$

$$\text{s.a.:} \quad \sum_j x_{ij} = 1, \qquad \forall i, \tag{6}$$

$$x_{ij} \leq y_j, \qquad \forall i, j \tag{7}$$

$$\sum_j y_j = p, \tag{8}$$

$$x_{ij}, y_j \in \{0, 1\}. \tag{9}$$

Obsérvese como la función objetivo (5) no incluye costes de apertura, debido a que el número de instalaciones son conocidas, así como su coste y por lo tanto no varía. Por lo tanto, se minimiza la distancia o coste de dar servicio. La restricción (6) obliga a que todo usuario debe ser servido por una instalación, la restricción de equilibrio (7) restringe que todo usuario debe ser abastecido por una planta abierta. Por último, la restricción (8) establece que el número total de plantas abiertas es igual al parámetro p fijado con anterioridad.

Una vez descrito el problema, los autores tratan de mostrar las técnicas heurísticas más destacadas recogidas en la literatura aplicadas al problema de la p-mediana, entre los que se destaca *Greedy, Stingy, Alternate y Composite Heuristics*.

La heurística *Greedy* comienza sobre un conjunto vacío de instalaciones abiertas, paso seguido se resuelve el problema de la *p-median* y se añaden las instalaciones una a una hasta llegar a un total igual al parámetro establecido p, seleccionando aquellas que reducen el coste o distancia total.

La heurística *Stingy*, al contrario de la heurística Greedy, comienza con todas las instalaciones abiertas eliminando una a una hasta alcanzar p, seleccionando aquellas que minimizan el coste en la función objetivo.

En cuanto al método denominado *Alternate*, las instalaciones se ubican en p puntos del conjunto L, en los que los usuarios son asignados a la instalación más cercana aplicando el problema *PMP* para los usuarios de cada instalación, una vez que no se produzcan cambios en los usuarios el algoritmo deja de iterar.

En último lugar, se describe el método conocido como *Composite Heuristics*. No puede considerarse un método en sí mismo, si no que hibrída varias técnicas heurístias. Así *Captivo (1991)*, propuso que en cada paso de Greedy se ejecuta el procedimiento Alternate.

Otro de ellos, la heurística de perturbación de *Salhi (1997)*, donde Stingy y Greedy se ejecutan uno tras otro, cada uno con un número determinado de pasos. Una de las técnicas más utilizada es la combinación entre Greedy e Interchange.

$$\begin{aligned} w_{ij} = & \sum_{u:c_1(u)\neq j} \max\{0, [d_1(u) - d(u,i)]\} \\ & - \sum_{u:c_1(u)=j} [\min\{d_2(u), d(u,i)\} - d_1(u)], \end{aligned} \tag{10}$$

Donde u,i y j, corresponden a los índices de un usuario y a las instalaciones de entrada y salida respectivamente; $d_1(u) = d(u, c_1(u))$ y $d_2(u)$ representan las distancias desde u a las instalaciones más cercanas y a la segunda más cercana, respectivamente. El primer sumatorio, referencia a los usuarios cuya instalación más cercana no es j, mientras que la segundas corresponden a los usuarios u reasignados a la nueva instalación i.

Uno de los estudios más importantes sobre esta temática corresponde a *Whitaker (1983)*, donde describe la llamada heurística de intercambio rápido. *Hansen y Mladenovi'c (1997)* la aplicaron como una subrutina de una heurística de *búsqueda de vecindario variable (VNS)*. Recientemente, *Resende y Wernek (2003)* sustituye la función anteriormente descrita por:

$w_{ij} = \sum_{u \in U} \max\{0, d_1(u) - d(0,1)\} - \sum_{u|c_1(u)=j}[d_2(u) - d_1(u)] + e_{ij}$.

Aquí, la primera suma representa el beneficio de insertar la instalación i, y la segunda recoge las pérdidas de la eliminación de la instalación j. En último lugar, e_{ij} proviene de una matriz que establece la clasificación de las distancias de cada usuario a todas las instalaciones potenciales.

Otra variante de la resolución de PMP mediante la búsqueda local de intercambio es la propuesta por *Kochetov y Alekseeva (2004)*,), denominada *LK (Lin-Keringham)*, el cual estableciendo un número k de movimientos de intercambios, busca la mejor solución de vecindad de 1 intercambio entre dos instalaciones, intercambiándolas para obtener una nueva solución y observar la posible mejora tras el intercambio.

En este apartado se hace un repaso de varias técnicas metaheurísticas desarrolladas para resolver el problema de la *p-median*, como: *Heurística lagrangiana, búsqueda Tabú (TS), búsqueda de vecindario variable (VNS), búsqueda genética, búsqueda de dispersión, GRASP con Path relinking, concentración heurística, redes neuronales, optimización colonia de hormigas.*

En la heurística lagrangiana, añade al modelo PMP multiplicadores lagrangianos u_i donde $i \in U$, así como la relajación de la restricción (6), por lo que la modificación heurística propone el siguiente modelo respecto al de *p-median*:

$$\max_{u} \min_{x,y} \sum_{i} (\sum_{j} d_{ij} - u_i) x_{ij} \tag{11}$$

s.a.: (7), (8) *y* (9)

Los pasos que sigue esta técnica pasan por establecer los valores de los multiplicadores u_i, encontrar los valores correspondientes de x_{ij} e y_i, y, por último, ajustar los multiplicadores. La relajación de la restricción (6), la cual establece que todo usuario debe ser servido por una instalación, puede ser alterada excepto por el hecho de que la asignación de un usuario es a su instalación abierta más cercana. Esta heurística es capaz de establecer cotas superiores e inferiores al problema, siendo la cota superior la mejor solución factible tras las iteraciones, y la cota inferior como el mayor valor de la nueva función objetivo (11).

Por último, la modificación de los multiplicadores de lagrange según los resultados obtenidos, proponen distintos enfoques como la optimización por subgradientes, la relajación semilagrangiana...

El método de *búsqueda Tabú (TS)*, para la resolución del problema *PMP* ha sido propuesto por distintos autores a lo largo del tiempo en los que se discute sobre el tamaño de las listas tabú, así como la vecindad de las posibles soluciones... En la literatura se recogen diversas variantes del método, *Mladenovi'c et al., (1995, 1996), Voss (1996), Glover y Laguna (1993), Salhi (2002)* que, centrado en la estructura de vecindad, propone que tras un movimiento de intercambio las instalaciones posibles de entrada se restringen a un total de K instalaciones más cercanas. *Kochetov (2001)* sugiere una búsqueda Tabu probabilística. Siendo $N(x)$ la vecindad de 1 intercambio de cualquier solución x, cuyo vecindario restringido $Nr(x) \subset N(x)$. Para el que cada $y \in N(x)$ se incluye en $Nr(x)$ en función a un umbral probabilístico propuesto por *Kochetov (2001)*.

La técnica búsqueda de *vecindario variable (VNS)* se encuentra contenida en varios estudios asociado al problema de *p-median*. En el primero, *(Hansen y Mladenovi'c, 1997)*, realiza un amplio análisis de varias estrategias, definiendo VNS como la técnica que, dada una solución inicial busca un óptimo a través del intercambio de soluciones entre las de su entorno. Los problemas de gran tamaño de *PMP*, estudiado en *Hansen et al. (2001)* aplica VNS como una variante de descomposición de VNS (VNDS), descomponiendo el problema VNS en varios subproblemas. Dos trabajos de *VNS Paralelo* para *PMP* son *García-López et al. (2002) y Crainic et al. (2004)*.

Por otra parte, para el *algoritmo genético (AG)* se proponen varias heurísticas como en *Hosage y Goodchild (1986)*, en la que mayormente penalizan el número de instalaciones abiertas, ofreciendo resultados deficientes incluso para problemas pequeños. Un mejor resultado fue ofrecido por *Dibbie y Densham (1993)*, en el que cada individuo tiene p genes, y cada gen representa un índice de facilidad. El tamaño del problema fue $n = m = 150 y p = 9$, consiguiendo unos resultados similares a la búsqueda local de intercambio, pero con un

mayor coste computacional. Por último, en *Alp et al. (2003)*, los resultados son mejores que las anteriores propuestas, donde la novedad radica en que se evita el operador de mutación y los nuevos miembros se generan utilizando operadores de selección y cruce. Prosigue seleccionando dos soluciones al azar y toman la unión de estas, en las que para poder ser viable se aplica la heurística clásica Stingy.

La metaheurística de *búsqueda dispersa (SS)* se inicializa seleccionando un conjunto de tamaño moderado entra una amplia población de soluciones, generando este conjunto de forma iterativa. Se combinan las soluciones del conjunto de referencia, aplicando un procedimiento de búsqueda local para mejorar la solución, repitiendo hasta alcanzar un criterio de parada. A diferencia de otras metaheurísticas, no proporciona una única solución sino, un conjunto reducido de soluciones. *García López (2003)* diseña un SS para el problema de la *p-median* en el que introduce una distancia en el espacio de soluciones, definiendo esta distancia entre J e I,

$$d(I,J) = \sum_{i \in I} \min_{j \in J} d_{ij} + \sum_{j \in J} \min_{i \in I} d_{ij}$$

El conjunto de referencia de soluciones se compone de k (un parámetro) mejores soluciones, combinando las soluciones del conjunto: primeramente encontrando el conjunto de instalaciones en cada solución del subconjunto; seguidamente, para encontrar el tamaño de p, se añaden nuevas instalaciones regido por reglas predefinidas; tras estos primeros pasos, las soluciones combinadas son mejoradas por una búsqueda local, incorporando la solución al conjunto de referencia mejorando alguna de sus k mejores soluciones.

GRASP con *Path relinking*, trata de una heurística en la que cada iteración se constituyen un número de puntos iniciales mediante un paso aleatorio *Greedy*, seguido de una búsqueda local entre dos soluciones de un conjunto de soluciones. Debido a que la distancia entre dos soluciones se modifica repetidamente antes de la búsqueda local, guarda similitud con la heurística *VNS*.

El método de *concentración heurística (HC)* se compone de dos etapas, la primera en la que se consiguen las soluciones siguiendo el algoritmo *Add-and-Drop*, q veces, reteniendo las m mejores dando lugar estas soluciones a un conjunto de concentración. En la segunda etapa, se limita el conjunto potencial de instalaciones a dicho conjunto de concentración y resuelve el modelo.

La *optimización por colonias de hormigas (ACO)*, basado en el comportamiento de las hormigas por su facilidad para encontrar el camino más corto, la idea principal de este método de optimización es el de utilizar la información generada en cada iteración y con esta conducir la búsqueda a la zona prometedora en el espacio de soluciones. Este método se recoge en *Lenova y Loresh (2004)*, como en [10], en el que inicialmente establece una solución aleatoria como todas las instalaciones potenciales en L; donde una instalación j que debe abandonarse se selecciona al azar del conjunto de vecindades de abandono.

Por último, *Domínguez Merino y Muñoz Pérez (2002)* proponen una red neuronal de dos capas para resolver el problema de la *p-median*.

En el artículo, los autores repasan y muestran las metodologías heurísticas y metaheurísticas con las que resolver u optimizar el problema de la *p-median*. Tras esta exposición, hace replantear al lector si la selección de métodos heurísticos y metaheurísticos mejoran o cambian el enfoque de este tipo de problema combinatorio, como el *PMP*. Se ha visto como en los casos donde el problema PMP ha sido pequeño, las distintas técnicas han logrado una solución óptima y de manera eficiente. El uso de la metaheurística ha permitido mejorar la calidad de las soluciones en instancias de gran tamaño. Estos enfoques, intensifican la búsqueda de regiones candidatas en el espacio de soluciones, reduciendo el tiempo computacional de resolución.

??caso real

Simultaneous feature selection and clustering using mixture models

Los autores *Martin H.C. Law, Student Member, Mario A.T. Figueiredo, Senior Member, y Anil K. Jain, Fellow*, abordan un problema a menudo poco estudiado como es el de la selección de características en el

clustering o análisis de conglomerados. A grandes rasgos pueden dividirse en: clustering jerárquico, el cuál va creando jerarquías a medida que itera sobre los datos generando los grupos, mientras que el clustering no jerárquico divide los datos en grupos disjuntos, es decir, no dependen jerárquicamente entre ellos (disjuntos).

Los algoritmos de clustering pueden tener como entrada de datos una matriz de proximidad que contiene similitudes/disimilitudes entre cada par de objetos, o una matriz de patrones en la que cada elemento se describe mediante un vector de características, siendo esta última tipología los datos de entrada.

Por ello, la selección de características es un proceso importante, con el fin de conseguir únicamente aquellas realmente determinantes. A lo largo del tiempo, la selección de variables ha sido un tema muy estudiado en los algoritmos de aprendizaje supervisado, como en clasificación de imágenes, biología molecular, etc. Sin embargo, respecto a los algoritmos de aprendizaje no supervisado no ha sido tal la literatura sobre esta problemática.

La necesidad de establecer un número k de grupos a priori en este tipo de algoritmos y la interrelación entre el número de grupos y el subconjunto de características, añade dificultad al problema. Los autores pretenden dar solución al problema de selección de variables en el aprendizaje no supervisado, con la propuesta de método de estimación en lugar de un método combinatorio. Para ello, proponen el uso del algoritmo *EM (Expectation-Maximization)* el cual proporciona una probabilidad de pertenencia de cada objeto a cada grupo formado. Conjuntamente para lograr discriminar correctamente entre las características relevantes e irrelevantes, el criterio *MML (longitud mínima mensaje)*[22] consigue que los valores de las características irrelevantes se acerquen a cero.

Acerca del tema, la mayoría de estudios versan sobre la selección de características en el aprendizaje supervisado como la regresión, por su parte, los algoritmos de selección de características se clasifican entre: filtro y envoltura.

Los métodos de selección de filtro se caracterizan por determinar la relación entre las variables y las etiquetas de clase mediante pesos, sin tener en cuenta el algoritmo utilizado posteriormente. Se consideran características irrelevantes aquellas que son condicionalmente independiente de las etiquetas dadas otras características.

Por otra parte, los métodos de envoltura usan el algoritmo en sí mismo para medir la importancia de las características en el conjunto de los datos, teniendo relevancia el algoritmo utilizado. La solución de ambos enfoques implica búsquedas combinatorias, por lo que en estructuras de datos de gran dimensionalidad el coste computacional es alto. Para resolverlo, se proponen técnicas heurísticas como búsquedas secuenciales (Forward selection, Backward selection), búsqueda genética, búsqueda bidireccional...

Lo descrito en este punto, solo tiene funcionalidad si las características tienen etiquetas de clase, determinando a su vez que, los métodos que no precisen de etiquetas de clase para conseguir un resultado pueden ser utilizado en el aprendizaje no supervisado. Existen algoritmos que asignan pesos a diferentes características para determinar su relevancia. En este estudio se propone el algoritmo *EM* basado en la contracción bayesiana para el aprendizaje no supervisado. A continuación, se propone un análisis clúster basado en modelos con selección de características abordado a través del algoritmo *EM (Expectation-Maximization)*. En este algoritmo, se calcula la probabilidad de que cada punto pertenezca a cada grupo, teniendo que cada grupo proviene de una distribución estadística probabilística. Realizándo el clustering aprendiendo los parámetros del modelo y las probabilidades asociadas.

La densidad de mezcla finita es definida como:

$$p(y) = \sum_{j=1}^{K} \alpha_j p(y|\theta_j)$$

Donde, $\forall j, \alpha_j \geq 0; \sum_j \alpha_j = 1$, cada θ_j es el conjunto de parámetros del componente j. La búsqueda de la estimación es inferir θ a partir de un conjunto de N datos $y = \{y_1, ..., y_N\}$, en el que j hace referencia al j-ésimo componente K, i al i-ésimo elemento de los N datos y l a la l-ésima característica.

Tras la definición de la función de densidad del algoritmo *EM*, este se basa en un conjunto de N etiquetas latentes $Z = \{z_1, \ldots, z_N\}$, donde $z_i = z_{il}, \ldots, z_{iK}$ con $z_{ij} = 1$ representa si y_i pertenece al clúster j.

El algoritmo genera una serie de valores para los parámetros θ, mediante una secuencia de dos pasos, el primer paso denominado como expectation trata de asignar cada elemento y_i a un cluster j, maximizando la esperanza de que cada cada uno de estos elementos pertenezca a un grupo, $W = E[z_{ij}|y, \hat{\theta}_j]$. El segundo paso, maximization actualiza el valor de los parámetros calculados en el primer paso para cada modelo, hasta que cada elemento y_i no cambie de clúster. En cuanto a la saliencia de las características, se definen como la forma de generar el valor de los parámetros de las características diferenciándolas entre relevantes e irrelevantes, derivados del algoritmo *EM (Maximization Expectation)*. Los autores han definido como variable irrelevante l cuando su distribución sea independiente de las etiquetas de clase, identificando $\phi = \{\phi_1, ..., \phi_D\}$ como un conjunto de parámetros binarios donde

$$\phi_l = \begin{cases} 1, \text{si la característica es relevante} \\ 0, \text{ en caso contrario} \end{cases}$$

Definiendo la saliencia como la probabilidad de que la característica l es relevante, $p_l = P(\phi_l = 1)$

$$p(y|\theta) = \sum_{j=1}^{K} \alpha_j \prod_{l=1}^{D} (p_l p(y_l|\theta_{jl}) + (1 - p_l) q(y_l|\lambda_l)), \tag{6}$$

donde $\theta = \{\{\alpha_j\}, \{\theta_{jl}\}, \{\lambda_l\}, \{p_l\}\}$, es el conjunto de parámetros del modelo.

El modelo *EM* para la estimación de parámetros contiene, Z (etiquetas de clase ocultas) y ϕ como variables latentes. Los investigadores añaden al paso *expectation* del algoritmo los siguientes cálculos:

$$a_{ijl} = P(\phi_l = 1, y_{il}|z_i = j) = p_l p(y_{il}|\theta_{jl}), \tag{7}$$

$$b_{ijl} = P(\phi_l = 0, y_{il}|z_i = j) = (1 - p_l) q(y_{il}|\lambda_l), \tag{8}$$

$$c_{ijl} = P(y_{il}|z_i = j) = a_{ijl} + b_{ijl}, \tag{9}$$

$$w_{ij} = P(z_i = j|y_i) = \frac{\alpha_j \prod_l c_{ijl}}{\sum_j \alpha_j \prod_l c_{ijl}}, \tag{10}$$

$$u_{ijl} = P(\phi_l = 1, z_i = j|y_i) = \frac{a_{ijl}}{c_{ijl} \cdot w_{ij}}, \tag{11}$$

$$v_{ijl} = P(\phi_l = 0, z_i = j|y_i) = w_{ij} - u_{ijl}, \tag{12}$$

Mientras que en el paso *maximization* los parámetros calculados son

$$\widehat{\alpha_j} = \frac{\sum_i w_{ij}}{\sum_{ij} w_{ij}} = \frac{\sum_i w_{ij}}{n}, \tag{13}$$

$$\widehat{\text{Mean in}}\ \theta_{jl} = \frac{\sum_i u_{ijl} y_{il}}{\sum_i u_{ijl}}, \tag{14}$$

$$\widehat{\text{Var in}}\ \theta_{jl} = \frac{\sum_i u_{ijl} (y_{il} - (\widehat{\text{Mean in}} \theta_{jl}))^2}{\sum_i u_{ijl}}, \tag{15}$$

$$\widehat{\text{Mean in}}\ \lambda_l = \frac{\sum_i (\sum_j v_{ijl}) y_{il}}{\sum_{ij} v_{ijl}}, \tag{16}$$

$$\widehat{\text{Var in}}\ \lambda_l = \frac{\sum_i (\sum_j v_{ijl}) (y_{il} - (\widehat{\text{Mean in}} \lambda_l))^2}{\sum_{ij} v_{ijl}}, \tag{17}$$

$$\widehat{p_l} = \frac{\sum_{ij} u_{ijl}}{\sum_{i,j} u_{ijl} + \sum_{i,j} v_{ijl}} = \frac{\sum_{ij} u_{ijl}}{n}, \tag{18}$$

La variable u_{ijl} representa la importancia del i-ésimo elemento para el j-ésimo componente cuando se utiliza la l-ésima variable. El término $\sum_j u_{ijl}$, puede interpretarse como la probabilidad de que ϕ_l es igual a uno.

Para conseguir un resultado óptimo se necesita establecer un número correcto de K grupos, por lo que una buena inicialización es esencial para conseguir un buen óptimo local. Para ello en el artículo se ha propuesto la inclusión de un criterio *MML (Maximum Message Length)* el cual minimiza respecto a θ.

El algoritmo EM subre una modificación en las ecuaciones (13), (18).

$$\widehat{\alpha_j} = \frac{\max\left(\sum_i w_{ij} - \frac{RD}{2}, 0\right)}{\sum_j \max\left(w_{ij} - \frac{RD}{2}, 0\right)}, \tag{21}$$

$$\widehat{p_l} = \frac{\max\left(\sum_{i,j} u_{ijl} - \frac{KR}{2}, 0\right)}{\max\left(\sum_{i,j} u_{ijl} - \frac{KR}{2}, 0\right) + \max\left(\sum_{i,j} v_{ijl} - \frac{S}{2}, 0\right)}, \tag{22}$$

Ambas ecuaciones en el paso *maximization*, consigue que determinadas α_j sean 0 y p_l sean 0 o 1, por lo que, cuando p_l llega a cero, la característica l deja de ser saliente y se eliminan p_l y $\theta_{1l}, \ldots, \theta_{Kl}$; cuando p_l llega a 1, se eliminan θ_l y p_l, consiguiendo de esta manera un buen criterio de selección, debido a la convergencia de α_j y p_l a 0. A continuación, tras haber calculado los anteriores términos, se calcula la saliencia de las características que mejor discriminan entre los distintos clúster, logrando maximizar la distancia entre los clúster, son definidos por una medida cuantitativa:

$$J = \sum_{i=1}^{N} logP(z_i = t_i | y_i$$

Donde $t_i = \arg\max_j P(z_i = j|y_i)$ siendo J la suma de los logaritmos de la probabilidad posterior de pertenencia.

En esta sección, se aplica el algoritmo propuesto por los autores en varios conjuntos de datos con distintas características. El primer set de datos de reconocimiento de vinos *(wine)*, contiene 178 observaciones, 13 variables y tres categorías. El dataset de diagnóstico de cáncer de mama de Wisconsin *(wdbc)* utilizado para detectar tumores malignos o benignos, contiene 30 características y 576 puntos de datos. El conjunto de segmentación de imágenes *(image)* contiene 2.320 elementos con 18 características de 7 categorías. El conjunto de datos textura *(texture)* consta de 4.000 observaciones, 19 dimensiones con 4 etiquetas de clase. Por último, el conjunto de datos *(zer)* está compuesto de 2.000 elementos representados por 47 características y 10 etiquetas de clase. Estos conjuntos de datos fueron recogidos para métodos de aprendizaje supervisado, por lo que en el experimento propuesto no se utilizan las etiquetas de clase, excepto para la evaluación de los resultados. Se han dividido los conjuntos de datos en dos: una parte de entrenamiento y la otra de prueba.

Se ha procedido a ejecutar el algoritmo *EM*, siendo post-procesados los valores de saliencia de las características según la medida cuantitativa J maximizando la distancia entre los clúster, anteriormente descrita. Una vez realizado el proceso, se evalúa la validez de los resultados de agrupación, con las etiquetas de clase existentes en cada uno de los conjuntos de datos. Asignando cada elemento del conjunto de datos al componente que probablemente lo haya generado.

A continuación, se calculan las tasas de error en los datos de prueba, comparando las mezclas gaussianas seleccionando directamente las características (sin saliencia) y creando un vector de valores sobre estas características (con saliencia). El algoritmo detallado en el artículo reduce las tasas de error comparando si se seleccionan todas las características. La mayor mejora se encuentra en el conjunto de datos imágenes, debido al alto número de componentes. En zernike el reconocimiento de imágenes de dígitos, es un problema muy complicado de resolver satisfactoriamente en el lenguaje no supervisado.

Respecto a la exigencia computacional, el algoritmo propuesto requiere de un número mayor de iteraciones respecto al algoritmo EM estándar debido al incremento del número de parámetros a calcular. Si bien es cierto que, el algoritmo propuesto es capaz de determinar tanto el número de clúster como los subconjuntos significativos de características, por lo que, en el caso de querer llegar al mismo objetivo con el algoritmo EM estándar, el tiempo de computación es mayor que la del algoritmo desarrollado en el artículo. Una de

las limitaciones existentes en el algoritmo es la suposición de independencia entre las variables. En el caso de no cumplirse esta condición, el resultado final puede verse alterado, no en cuanto a la calidad de los clúster o a la precisión de un clasificador, sino a la selección de características relevantes. Esta problemática se resuelve con el uso del post procesamiento J propuesto anteriormente, ya que tiene en cuenta la distribución posterior, descartando las características que no ayudan a identificar los clúster.

Ocurre que el investigador o la persona correspondiente a realizar el análisis, tiene cierto conocimiento de los datos y puede conocer las etiquetas de clase de distintos componentes gaussianos. Este hecho ayuda a supervisar el resultado del algoritmo. Los autores han presentado un algoritmo EM capaz de estimar la importancia de las características y el mejor número de componentes para la agrupación de mezclas gaussianas en el artículo. Finalmente, como líneas futuras o modificaciones del algoritmo que se ha propuesto, es la de reducir el tiempo de ejecución cuando el tamaño de los datos es muy grande, también intentar modelar el caso de dependencia de las características entre sí.

Existen diversas áreas de aplicación del algoritmo EM [21]como método de clasificación de objetos a grupos. Uno de los casos de aplicación reciente podría ser el de medir la gravedad de un paciente infectado por COVID-19. En la recogida de datos se puede etiquetar en función a gente enfermada con anterioridad entre, hospitalización UCI o atención médica, y recogiendo diversas variables que ayuden a su identificación. El algoritmo EM se puede encuadrar como un tipo de clúster probabilístico, es decir, proporciona el valor de la probabilidad de pertenecer a cada grupo. La etiquetación de los sujetos no tiene utilidad en el algoritmo salvo para validar los resultados.Por lo tanto, dicho algoritmo agrupará los sujetos según las variables relevantes proporcionadas por los técnicos.

En función a los programas de análisis estudiados en el grado, destacando para implementar este problema LINGO, debido a ser el utilizado en asignaturas de modelización de problemas combinatorios.

A Solution Proposal for the Capacitated P-Median Problem with Tabu Search

En el artículo "A Solution Proposal for the Capacitated P-Median Problem with Tabu Search" los autores *Mauricio Romero Montoya, María Beatríz Bernábe Loranca, Rogelio González Velázquez, José Luis Martínez Flores, Horacio Bautista Santos, Abraham Sánchez Flores, Francisco Macias Santiesteban*, tratan de resolver un problema de localización óptima de p puntos a través de la búsqueda tabú (TS).

En los problemas de localización de instalaciones, consta de elegir la ubicación óptima de un conjunto de localizaciones, bajo un conjunto de restricciones. Entre los numerosos métodos de resolución de este problema, son característicos por su eficacia, el problema de la *p-median*.

El problema de la *p-median capacitado (CPMP)* difiere del modelo no capacitado en la limitación de capacidad de cada una de las instalaciones. Se plantea el uso de la *Búsqueda Tabu (TS)* para conseguir una aproximación más eficiente al punto óptimo. Las distintas suposiciones a la hora de afrontar un problema de localización, como la capacidad de abastecimiento de las plantas, costes de apertura de plantas, coste de cambio de ubicación, preferencias de los clientes... hace que no exista un modelo genérico que pueda afrontar diversos problemas.

A lo largo de la historia, desde Euclides y Pitágoras, pasando por el emperador Constantino resolviendo la ubicación correcta de tropas en distintos lugares, Fermat hasta Sylvester en el siglo XIX han abordado esta problemática.

A raíz de los múltiples estudios sobre la problemática de la localización óptima, *Hakimi (1964)* propone el problema *p-median*. Se trata de un problema NP, por lo que el tiempo computacional de resolución incrementa exponencialmente al añadir nuevas variables al problema, por lo que se complementa con técnicas heurísticas las que consiguen solucionar este caso, entre otros la relajación de Lagrange o Branch-and-Bound como los más estudiados en este campo.

Por lo tanto, se tiene que, el problema *p-median* general, considera que las instalaciones son capaces de abastecer a todos los clientes sin poner en duda su capacidad. Sin embargo, las capacidades finitas de estas provocan que puedan considerarse distintas situaciones, como por ejemplo que un cliente no pueda ser

abastecido por su instalación más cercana. Todo ello, da lugar a una variante del problema inicial, *p-median capacitado (CPMP)* el cual añade una restricción de capacidad sobre las instalaciones abiertas añadiendo esto una mayor dificultad a su resolución.

Mediante técnicas heurísticas o metaheurísticas se intenta llegar a la solución, dos artículos de interés por la sociedad científica, como *Sorensen (2008)* basado en resolución del problema *CPMP* por un procedimiento de búsqueda Tabu (TS) mejorando cualquier trabajo anterior sobre CPMP, respaldando las conclusiones del artículo Arostegui et al. (2006). El problema de la p-median capacitado se describe como, dado un conjunto de n clientes, con una demanda conocida para cada uno de ellos, se busca encontrar p medianas, asignando cada cliente a una determinada mediana minimizando la distancia entre los clientes y medianas, sin poder ser excedido su capacidad. Matemáticamente el problema se formula:

$$z = \min \sum_{i \in N} \sum_{j \in N} d_{ij} x_{ij}, \tag{1}$$

$$\text{s.a.:} \quad \sum_{j \in N} x_{ij} = 1, \qquad \forall i \in N, \tag{2}$$

$$\sum_{j \in N} x_{jj} = p, \tag{3}$$

$$\sum_{i \in N} q_i x_{ij \leq Q x_{jj}}, \qquad \forall j \in N, \tag{4}$$

$$x_{ij} \in \{0, 1\}, \qquad \forall i \in N, \forall j \in N, \tag{5}$$

Donde $N = 1, ..., n$ es el conjunto de clientes i, siendo p el número de medianas; q_i representa la demanda de cada cliente i y Q la capacidad que tiene cada instalación para abastecer a los puntos i; d_{ij} corresponde a la matriz de distancias entre el cliente i y la instalación j, x_{ij} la matriz de asignación, teniendo que $x_{ij} = 1$ si el cliente i es asignado a la mediana j, mientras que $x_{jj} = 1$ determina si se selecciona la mediana j, 0 en caso contrario.

En dicho problema, la función objetivo (1) minimiza la distancia entre el cliente i y la instalación j. En (2) se garantiza la asignación de cada cliente a una instalación, la tercera restricción establece el número de medianas a seleccionar, mientras que la restricción (4) hace referencia a la capacidad de la mediana, teniendo que la suma de la de la demanda de los clientes sea menor que la capacidad de las medianas asignadas.

Los autores han utilizado la técnica *search clustering (CS)*, la cual trata de encontrar áreas prometedoras de ser seleccionadas en el espacio definidas como clústers, donde dichos clúster son definidos como $G = c; r; s$ donde c es el centro del clúster, r el radio del área, estableciendo la distancia máxima o mínima, desde el centro del clúster y s la estrategia de búsqueda. De forma simultánea al *search clustering (CS)* se aplica el algoritmo de *búsqueda Tabu (TS)*, generando una aproximación a posibles candidatos a óptimo entre el total de soluciones posibles. Dicho algoritmo, dada una solución inicial de puntos candidatos, define un vecindario de soluciones evaluando cada punto y seleccionando aquellos que minimizan o maximizan el criterio deseado. La dotación de inteligencia del algoritmo radica en la posibilidad de guardar dichos resultados en memoria, es decir, memoriza aquellos cambios entre los puntos que no son candidatos a soluciones (en función al criterio establecido; minimizar o maximizar...). Los puntos con peor valor del criterio propuesto no serán evaluados en posteriores iteraciones, generando la denominada Lista Tabú (LS).

En función a los resultados obtenidos por los investigadores en la aplicación del algoritmo de *búsqueda Tabu*, se determina la eficiencia de este método para la selección de conjuntos de soluciones factibles en el problema de la *p-median capacitada (CPMP)*. Como trabajos futuros con esta técnica sobre esta problemática, se pretende desarrollar una serie de aplicaciones en cadenas de suministros buscando seleccionar correctamente el emplazamiento de instalaciones.

Se propone un problema referente a la selección de técnicos para la tasación de un número de viviendas. Para ello y en función al problema de la *p-median* capacitada se determinan los siguientes parámetros. Suponiendo que el problema se encuadra dentro de la Comunidad Valenciana, seleccionamos el número de clientes como

los municipios que tienen viviendas para ser tasadas; referente a las localizaciones, lo asignamos como el número de técnicos potenciales a poder tasar dichas viviendas, estableciendo como capacidad de estos el número máximo de viviendas a tasar y la demanda de cada 'cliente' al número de viviendas en ese municipio.

Las distancias calculadas entre 'clientes' e 'instalaciones', será la distancia entre el lugar de residencia del tasador, ya que es el punto desde donde se desplaza cuando tiene que tasar una vivienda, a el municipio donde se encuentra la vivienda a tasar. Siguiendo la notación anterior, donde $N = \{1, \ldots, n\}$ es el conjunto de clientes a asignar y técnicos.

- q_i = número total de viviendas en el municipio i, $i = 1, ..., n$.
- Q = número total de viviendas que puede tasar cada técnico.
- d_{ij} = distancia desde el municipio i hasta el técnico $j, \forall i \in N, \forall j \in m$.

Cuyas variables de decisión son:

$$x_{jj} = \begin{cases} 1, \text{si un técnico } j \text{ es asignado para realizar una tasación}, \forall j \in N \\ 0, \text{ en caso contrario} \end{cases}$$

$$x_{ij} = \begin{cases} 1, \text{ si la vivienda } i \text{ es tasada por el técnico } j, \forall i \in N, \forall j \in N \\ 0, \text{ en caso contrario} \end{cases}$$

$$z = \min \sum_{i \in N} \sum_{j \in N} d_{ij} x_{ij}, \tag{1}$$

$$\text{s.a.:} \quad \sum_{j \in N} x_{ij} = 1, \qquad \forall i \in N, \tag{2}$$

$$\sum_{j \in N} x_{jj} = p, \tag{3}$$

$$\sum_{i \in N} q_i x_{ij \leq Q x_{jj}}, \qquad \forall j \in N, \tag{4}$$

$$x_{ij} \in \{0, 1\}, \qquad \forall i \in N, \forall j \in N, \tag{5}$$

Por lo tanto, la función objetivo minimiza la distancia recorrida por el técnico a la hora de realizar una tasación, la restricción (3) propone que el número total de técnicos disponibles es el mismo que el p propuesto a priori. Diferenciándose del modelo *p-median* clásico, en (4) se establece que la demanda de cada municipio (n º total de viviendas a tasar) no puede superar el número de viviendas que puede tasar un técnico activo.

Conclusiones

Con este trabajo realizado, se ha expuesto los distintos tipos de análisis clúster en la actualidad, así como sus características y particularidades. Todo ello en función de los datos disponibles, o la finalidad del estudio llevado a cabo por el investigador. Se pretende con esto, abordar los distintos problemas en los que se pueda aplicar estos análisis.

En la segunda parte del trabajo, se pretende dar solución a la localización de plantas, como a la selección de variables relevantes en un análisis de datos, mediante el problema *p-median*. En la revisión realizada, se ahonda en distintos métodos heurísticos con los que resolver el problema *p-median* de forma óptima con tiempos de computación aceptables. Se recoge desde la localización de plantas de manera que se satisfaga la demanda de los clientes de manera óptima, hasta la selección de variables significativas y evitar con ello resultados erróneos debido a la inclusión de variables irrelevantes.

El interés de este trabajo radica en la novedad de dar solución al aumento de problemas relacionados con la localización de instalaciones y ampliar la capacidad de resolución de los mismos.

En los distintos problemas planteados de la *p-median*, se presupone que en un marco teórico, todas las instalaciones posibles de ser abiertas, siempre se encuentran disponibles para su uso, por lo que una posible linea futura de investigación sería el modelizar posibles fallos, o momentos en los que las instalaciones no se encuentren operativas.

En relación a lo estudiado en el grado de Estadística Empresarial, la asignatura cursada en cuarto curso *Gestión y Planificación de la Producción* en la que se aprende a modelizar situaciones en el ámbito de la empresa a través de modelos lineales, así como a su resolución mediante programas informáticos. La asignatura *Minería de Datos* de tercero, se estudiaron distintas técnicas multivariantes aplicada a los datos, en la que dentro de la materia se pudo aprender acerca del análisis clúster.

De manera indirecta, se consigue extraer conocimiento aplicable al caso en concreto de muchas asignaturas del grado, las cuales proporcionan solidez infórmatica y matemática, así como compresión de lo leído, entre otras, *Gestión de Carteras e Inversiones*,*Modelos de Optimización*...

Bibliografía

[1] Hubert, L., y Arabie, P. (1985), "Comparing partitions", Journal of Classification 2: 193-218.

[2] Rousseeuw, P. J. (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics 20: 53-65.

[3] Benati, Stefano, Garcia Quiles, Sergio & Puerto, Justo. (2018). "Mixed Integer Linear Programming and Heuristic Methods for Feature Selection in Clustering". Journal of the Operational Research Society. 69.

[4] Koehn, Hans & Steinley, Douglas & Brusco, Michael. (2010). "The p-Median Model as a Tool for Clustering Psychological Data". Psychological methods. 15. 87-95.

[5] Benati, Stefano & García Quiles, Sergio. (2012). "A p-median problem with distance selection".

[6] Payman Kaveh, Ali Sabzevari Zadeh and Rashed Sahraeian. (2010) "Solving Capacitated P-median Problem by Hybrid K-means Clustering and FNS Algorithm", International Journal of Innovation, Management and Technology vol. 1, no. 4, pp. 405-410.

[7] Nenad Mladenović, Jack Brimberg, Pierre Hansen, José A. Moreno-Pérez. (2007) "The p-median problem: A survey of metaheuristic approaches", European Journal of Operational Research, Volume 179, Issue 3, Pages 927-939.

[8] Law, Martin & Figueiredo, Mário & Jain, Anil. (2004). "Simultaneous feature selection and clustering using mixture models". IEEE transactions on pattern analysis and machine intelligence. 26. 1154-66.

[9] Romero Montoya, Mauricio & Bernábe Loranca, María Beatríz & González Velázquez, Rogelio & Flores, José Luis & Bautista-Santos, Horacio & Flores, Abraham & Santiesteban, Francisco. (2018). "A Solution Proposal for the Capacitated P-Median Problem with Tabu Search". Research in Computing Science.

[10] Muñoz B., César Adrián, Ramón Alfonso, Toro O., Eliana M.. (2011), "Comparación del desempeño del algoritmo genético de Chu-Beasley y el algoritmo colonia de hormigas en el problema de la p-mediana". Scientia Et Technica. XVII(47), 213-218.

[11] Universidad de Granada (2012). Medidas de Asociación, http://www.ugr.es/ gallardo/pdf/cluster-2.pdf.

[12] Universidad de Granada (2012). Métodos Jerárquicos de Análisis Cluster., http://www.ugr.es/ gallardo/pdf/cluster-3.pdf.

[13] Universidad de Granada (2012). Métodos no Jerárquicos de Análisis Cluster., http://www.ugr.es/ gallardo/pdf/cluster-4.pdf.

[14] Jain, A. K., & Dubes, R. C. (1988). "Algorithms for Clustering Data". Upper Saddle River, NJ: Prentice-Hall, Inc.

[15] Brusco, Michael & Koehn, Hans. (2008). "Optimal Partitioning of a Data Set Based on the p-Median Model". Psychometrika. 73. 89-105.

[16] Fernández Santana, Oscar. (1991). "The Cluster Analysis: application, Interpretation and Validation". Papers: revista de sociología, Nº 37, 1991 (Ejemplar dedicado a: El análisis multivariable de datos), págs. 65-76

[17] Hubert, Lawrence & Arabie, Phipps & Meulman, Jacqueline. (2001). "Combinatorial data analysis. Optimization by dynamic programming".

[18] J. Brusco, Michael & Stahl, Stephanie. (2005). "Branch-and-Bound Applications in Combinatorial Data Analysis".

[19] Scharl, Theresa & Leisch, Friedrich. (2006). "The stochastic QT-clust algorithm: evaluation of stability and variance on time-course microarray data". Compstat 2006–Proceedings in Computational Statistics.

[20] Bornstein, Claudio & Campêlo, Manoel. (2004). "An ADD/DROP procedure for the capacitated plant location problem". Pesquisa Operacional. 24. 151-162.

[21] Bruijns, Jan. (2009). "Acceleration of the Expectation-maximization Algorithm for a Twofold Gaussian Mixture Model by using the Histogram of the Observations Instead of the Observations - Evaluation of its Accuracy by Generated Histograms". VISAPP 2009 - Proceedings of the 4th International Conference on Computer Vision Theory and Applications.

[22] Silvestre, C. & Cardoso, M. & Figueiredo, M. A. T. (2016). "Model selection in discrete clustering: The EM-MML algorithm", Proc ERCIM - International Conference of the ERCIM Working Group on Computational and Methodological Statistics CMStatistics, Seville, Spain, Vol. N/A, pp. 200 - 200.

Anexo

p	Global Optimum	Reduction p to p+1 in %	Silhouette Index (SI_k)
2	2871	-	.1630
3	2385	16.93	.2227
4	1946	18.41	.3129
5	1619	16.80	.3843
6	1358	16.12	.4330
7	1119	17.60	.4884
8	964	13.85	.5514
9	885	8.20	.5535
10	817	7.68	.5580
11	751	8.08	.5594
12	696	7.32	.5454
13	641	7.90	.5292
14	591	7.80	.5196

Tabla 1: Soluciones p-median globalmente óptimas para $2 \leq p \leq 14$,porcentaje de reducción de la función objetivo e índices de silueta

Cluste	**1**	**2**	**3**	**4**
	Apple (1)	**Broccoli (7)**	**Bagel (14)**	**Pretzels (22)**
	Watermelon (2)	Lettuce (6)	Rice (12)	Crackers (20)
	Orange (3)	Carrots (8)	Bread (13)	Popcorn (23)
	Banana (4)	Corn (9)	Oatmeal (15)	Nuts (24)
	Pineapple (5)	Onions (10)	Cereal (16)	Potato Chip (25)
		Potato (11)	Muffin (17)	
			Pancake (18)	
			Spaghetti (19)	
			Granola Bar (21)	
Cluster	**5**	**6**	**7**	**8**
	Pie (30)	**Cheese (35)**	**Water (38)**	**Pork (42)**
	Doughnuts (26)	Yogurt (33)	Soda (39)	Hamburguer (40)
	Cookies (27)	Butter (34)		Steak (41)
	Cake (28)	Eggs (36)		Chicken (43)
	Chocolate Bar (29)	Milk (37)		Lobster (44)
	Pizza (31)			Salmon (45)
	Ice Cream (32)			

Tabla 2: Agrupación de alimentos globalmente óptima, con $p = 8$ medianas

				q-vars			Add-Drop		ILP-P1			ILP-P2	
Problem Name	**q***	**fo**	**it-best**	**time-best**	**fo**	**it-best**	**time-best**	**fo**	**root LP**	**time**	**fo or best UB**	**root LP**	**time**
MSA-A-15x4x50	20	43.89	1	0	43.89	9	0	43.89	40.95	0	43.89	43.89	0
MSA-B-15x4x50	20	41.44	1	0	41.44	9	0	41.44	38.62	0	41.44	41.44	0
MSA-C-15x4x50	20	46.14	1	0	46.14	9	0	46.14	42.22	0	46.14	46.14	0
MSA-D-15x4x50	20	45.47	1	0	45.47	9	0	45.47	42.03	0	45.47	45.47	0
MSA-E-15x4x50	20	39.61	1	0	39.61	9	0	39.61	36.59	0	39.61	39.61	0
MSA-A-15x4x100	20	43.63	2	0	43.63	17	0	43.63	39.96	0	43.63	43.63	0
MSA-B-15x4x100	20	43.61	1	0	43.61	17	0	43.61	39.72	0	43.61	43.61	0
MSA-C-15x4x100	20	40.77	1	0	40.77	17	0	40.77	38.53	0	40.77	40.77	0
MSA-D-15x4x100	20	39.57	1	0	39.57	17	0	39.57	37.46	0	39.57	39.57	0
MSA-E-15x4x100	20	46.74	2	0	46.74	17	0	46.74	43.51	0	46.74	46.74	0
MSA-A-15x4x500	20	43.38	1	0	43.38	19	0	43.38	39.38	0	43.38	43.38	0
MSA-B-15x4x500	20	47.17	2	0	47.17	19	0	47.17	44.17	0	47.17	47.17	0
MSA-C-15x4x500	20	45.05	2	0	45.05	19	0	45.05	40.91	0	45.05	45.05	0
MSA-D-15x4x500	20	43.46	2	0	43.46	19	0	43.46	39.89	0	43.46	43.46	0
MSA-E-15x4x500	20	46.60	1	0	46.60	19	0	46.60	42.14	0	46.60	46.60	0
MSA-A-15x4x1000	20	44.34	2	0	44.34	20	0	44.34	40.21	2	44.34	44.34	0
MSA-B-15x4x1000	20	42.01	2	0	42.01	20	0	42.01	38.44	1	42.01	42.01	0
MSA-C-15x4x1000	20	41.25	2	0	41.25	20	0	41.25	38.76	1	41.25	41.25	0
MSA-D-15x4x1000	20	39.41	2	0	39.41	20	0	39.41	36.82	1	39.41	39.41	0
MSA-E-15x4x1000	20	38.59	2	0	38.59	20	0	38.59	36.58	1	38.59	38.59	0
Average	**20**	**43.11**			**43.11**			**43.11**	**39.85**		**43.11**	**43.11**	
MSA-A-15x4x50	10	18.12	1	0	21.23	547	0	18.12	17.44	0	18.12	18.12	0
MSA-B-15x4x50	10	17.18	1	0	18.05	580	0	17.18	15.77	0	17.18	17.18	0
MSA-C-15x4x50	10	18.94	1	0	20.23	1876	0	18.94	17.39	0	18.94	18.94	0
MSA-D-15x4x50	10	18.89	1	0	20.36	13	0	18.89	18.08	0	18.89	18.89	0
MSA-E-15x4x50	10	15.55	1	0	17.49	34	0	15.55	14.97	0	15.55	15.55	0
MSA-A-15x4x100	10	16.87	2	0	18.10	378	0	16.87	15.68	0	16.87	16.87	0
MSA-B-15x4x100	10	17.48	2	0	19.17	892	0	17.48	17.28	0	17.48	17.48	0
MSA-C-15x4x100	10	16.75	2	0	18.42	333	0	16.75	15.68	0	16.75	16.75	0
MSA-D-15x4x100	10	16.33	2	0	17.82	201	0	16.33	16.21	0	16.33	16.33	0
MSA-E-15x4x100	10	19.53	2	0	21.18	686	0	19.53	18.80	0	19.53	19.53	0
MSA-A-15x4x500	10	16.53	2	0	21.39	1154	0	16.53	15.85	0	16.53	16.53	0
MSA-B-15x4x500	10	19.15	2	0	21.87	994	0	19.15	18.70	0	19.15	19.15	0
MSA-C-15x4x500	10	18.16	2	0	19.94	530	0	18.16	17.12	0	18.16	18.16	0
MSA-D-15x4x500	10	17.96	2	0	20.42	226	0	17.96	16.74	0	17.96	17.96	0
MSA-E-15x4x500	10	19.69	2	0	23.35	540	0	19.69	18.25	0	19.69	19.69	0
MSA-A-15x4x1000	10	17.26	2	0	19.22	550	0	17.26	16.90	0	17.26	17.26	0
MSA-B-15x4x1000	10	18.01	2	0	19.24	765	0	18.01	16.46	1	18.01	18.01	0
MSA-C-15x4x1000	10	16.32	2	0	20.42	1010	0	16.32	15.89	0	16.32	16.32	0
MSA-D-15x4x1000	10	16.16	2	0	18.36	11	0	16.16	15.36	0	16.16	16.16	0
MSA-E-15x4x1000	10	17.08	2	0	17.98	600	0	17.08	16.29	1	17.08	17.08	0
Average		**17.60**			**19.71**			**17.60**	**16.74**		**17.60**	**17.60**	
MSA-A-15x4x50	40	247.86	1	0	247.86	634	1	247.86	154.60	0	247.86	247.86	0
MSA-B-15x4x50	40	235.34	1	0	235.34	180	0	235.34	146.05	0	235.34	235.34	0
MSA-C-15x4x50	40	247.30	1	0	247.30	520	1	247.30	159.87	0	247.30	247.30	0
MSA-D-15x4x50	40	241.20	1	0	241.20	980	1	241.20	149.30	0	241.20	241.20	0
MSA-E-15x4x50	40	239.69	1	0	239.69	308	0	239.69	149.89	0	239.69	239.69	0
MSA-A-15x4x100	40	230.18	1	0	238.29	384	1	230.18	139.79	2	230.18	230.18	0
MSA-B-15x4x100	40	226.71	1	0	236.00	102	0	226.71	140.97	1	226.71	226.71	0
MSA-C-15x4x100	40	221.81	1	0	227.80	273	0	221.81	138.54	1	221.81	221.81	0
MSA-D-15x4x100	40	217.69	1	0	225.45	2773	4	217.69	135.03	1	217.69	217.69	0
MSA-E-15x4x100	40	236.60	1	0	245.25	553	1	236.60	145.58	2	236.60	236.60	0
MSA-A-15x4x500	40	185.61	1	0	226.71	164	0	185.61	118.34	7	185.61	170.83	167
MSA-B-15x4x500	40	204.59	1	0	234.22	801	1	204.59	126.77	15	204.59	176.45	964
MSA-C-15x4x500	40	189.23	1	0	228.33	706	1	189.23	124.20	9	189.23	173.99	107
MSA-D-15x4x500	40	201.97	1	0	227.99	2069	4	201.97	116.75	15	201.97	171.78	1158
MSA-E-15x4x500	40	205.43	1	0	230.80	343	1	205.43	123.63	16	205.43	177.57	772
MSA-A-15x4x1000	40	187.79	2	0	229.06	198	0	187.79	115.45	38	187.79	159.05	1904
MSA-B-15x4x1000	40	183.91	1	0	222.83	914	2	183.91	109.07	30	184.03	156.41	7200
MSA-C-15x4x1000	40	180.67	1	0	220.27	77	0	180.67	108.39	49	180.67	154.69	2867
MSA-D-15x4x1000	40	181.34	1	0	222.50	847	2	181.34	111.17	48	181.34	153.37	4797
MSA-E-15x4x1000	40	173.76	1	0	219.92	186	0	173.76	11.82	21	173.76	153.49	988
Average		**211.93**			**232.34**			**211.93**	**126.26**		**211.94**	**199.60**	

Tabla 3: Resultados computacionales: Aplicación a la reducción de dimensiones.

			q-vars				Add-Drop			ILP-P1			ILP-P2		
Problem Name	**q**	**fo**	**opt-it**	**time-opt**	**accuracy**	**fo**	**opt-it**	**time-opt**	**accuracy**	**fo**	**root LP**	**time**	**fo**	**root LP**	**time**
PRV-A-100x2x40	10	119.14	1	0	3	119.78	133	0	2	119.14	117.79	0	119.14	119.14	0
PRV-B-100x2x40	10	130.29	1	0	2	130.29	569	0	2	130.29	126.00	0	130.29	130.29	0
PRV-C-100x2x40	10	128.01	1	0	0	128.01	6	0	0	128.01	124.37	0	128.01	128.01	0
PRV-D-100x2x40	10	121.92	1	0	1	121.92	7	0	1	121.92	118.84	0	121.92	121.92	0
PRV-E-100x2x40	10	120.67	1	0	0	120.67	6	0	0	120.67	120.48	0	120.67	120.67	0
PRV-A-100x2x40	20	357.23	1	0	1	357.23	11	0	1	357.23	324.51	0	357.23	357.23	0
PRV-B-100x2x40	20	345.93	1	0	1	345.93	11	0	1	345.93	325.88	0	345.93	345.93	0
PRV-C-100x2x40	20	352.89	1	0	1	352.89	88	0	1	352.89	322.01	0	352.89	352.89	0
PRV-D-100x2x40	20	334.82	1	0	1	334.82	11	0	1	334.82	308.16	0	334.82	334.82	0
PRV-E-100x2x40	20	357.06	1	0	0	357.06	11	0	0	357.06	331.69	0	357.06	357.06	0
PRV-A-100x2x40	30	827.52	1	0	2	827.52	25	0	2	827.52	641.14	1	827.52	827.52	0
PRV-B-100x2x40	30	782.85	1	0	0	782.85	7	0	0	782.85	637.04	1	782.85	782.85	0
PRV-C-100x2x40	30	799.33	1	0	0	799.33	7	0	0	799.33	628.64	1	799.33	799.33	0
PRV-D-100x2x40	30	787.40	1	0	0	787.40	7	0	0	787.40	624.59	1	787.40	787.40	0
PRV-E-100x2x40	30	818.74	1	0	0	818.74	7	0	0	818.74	643.29	1	818.74	818.74	0
Average		**425.59**				**425.63**				**425.59**	**359.63**		**425.59**	**425.59**	
PRV-A-100x2x80	20	263.31	1	0	2	263.71	80	0	1	263.31	257.02	0	263.31	263.31	0
PRV-B-100x2x80	20	266.87	1	0	1	266.87	168	0	1	266.87	260.95	0	266.87	266.87	0
PRV-C-100x2x80	20	255.95	1	0	2	255.95	14	0	2	255.95	252.64	0	255.95	255.95	0
PRV-D-100x2x80	20	263.48	1	0	2	263.48	216	0	2	263.48	256.38	0	263.48	263.48	0
PRV-E-100x2x80	20	245.06	1	0	2	246.80	34	0	1	245.06	239.28	0	245.06	245.06	0
PRV-A-100x2x80	40	727.85	1	0	1	727.85	24	0	1	727.85	665.12	1	727.85	727.85	0
PRV-B-100x2x80	40	719.68	1	0	2	719.68	92	1	2	719.68	666.79	1	719.68	719.68	0
PRV-C-100x2x80	40	729.66	1	0	1	729.66	23	0	1	729.66	676.67	1	729.66	729.66	0
PRV-D-100x2x80	40	726.86	1	0	1	726.86	129	1	1	726.86	652.20	1	726.86	726.86	0
PRV-E-100x2x80	40	690.03	1	0	1	690.03	128	1	1	690.03	634.69	0	690.03	690.03	0
PRV-A-100x2x80	60	1694.27	1	0	1	1694.27	279	4	1	1694.27	1311.62	5	1694.27	1694.27	0
PRV-B-100x2x80	60	1627.80	1	0	1	1627.80	15	0	1	1627.80	1297.12	5	1627.80	1627.80	0
PRV-C-100x2x80	60	1654.34	1	0	0	1654.34	15	0	0	1654.34	1319.34	4	1654.34	1654.34	0
PRV-D-100x2x80	60	1635.18	1	0	0	1635.18	15	0	0	1635.18	1289.34	4	1635.18	1635.18	0
PRV-E-100x2x80	60	1621.49	1	0	2	1621.49	31	0	2	1621.49	1279.15	9	1621.49	1621.49	0
Average		**874.79**				**874.93**				**874.79**	**737.22**		**874.79**	**874.79**	
PRV-A-100x2x120	30	390.01	1	0	4	391.07	471	2	3	390.01	381.819	0	390.01	390.01	0
PRV-B-100x2x120	30	386.92	1	0	5	388.78	1556	5	4	386.92	377.796	1	386.92	386.92	0
PRV-C-100x2x120	30	386.30	1	0	3	386.30	371	2	3	386.30	380.574	1	386.30	386.30	0
PRV-D-100x2x120	30	376.49	1	0	2	376.49	375	1	2	376.49	369.607	0	376.49	376.49	0
PRV-E-100x2x120	30	370.08	1	0	4	374.20	841	3	2	370.08	362.519	0	370.08	370.08	0
PRV-A-100x2x120	60	1040.36	1	0	3	1040.36	215	3	3	1040.36	970.97	1	1040.36	1040.36	0
PRV-B-100x2x120	60	1046.18	1	0	0	1046.18	26	0	0	1046.18	970.81	1	1046.18	1046.18	0
PRV-C-100x2x120	60	1085.35	1	0	2	1087.07	357	5	2	1085.35	990.176	1	1085.35	1085.35	0
PRV-D-100x2x120	60	1067.89	1	0	1	1070.13	206	3	2	1067.89	989.424	2	1067.89	1067.89	0
PRV-E-100x2x120	60	1025.97	1	0	1	1025.97	26	0	1	1025.97	964.7	1	1025.97	1025.97	0
PRV-A-100x2x120	90	2391.07	1	0	2	2391.07	19	1	2	2391.07	1904.8	11	2391.07	2391.07	0
PRV-B-100x2x120	90	2366.39	1	0	1	2366.39	44	1	1	2366.39	1888.17	6	2366.39	2366.39	0
PRV-C-100x2x120	90	2469.22	1	0	3	2469.22	19	1	3	2469.22	1941.87	12	2469.22	2469.22	0
PRV-D-100x2x120	90	2451.00	1	0	2	2451.00	240	8	2	2451.00	1929.39	11	2451.00	2451.00	0
PRV-E-100x2x120	90	2340.39	1	0	1	2340.39	20	6	1	2340.39	1886.16	10	2340.39	2340.39	0
Average		**1279.57**				**1280.31**				**1279.57**	**1087.25**		**1279.57**	**1279.57**	

Tabla 4: Resultados computacionales: Selección de variables para datos sin estructura

			q-vars			Add-Drop			ILP-P1		ILP-P2	
Problem Name	**q**	**fo**	**opt-it**	**time-opt**	**fo**	**opt-it**	**time-opt**	**obj/best UB**	**root LP**	**obj/best UB**	**root LP**
WRN-A-96x8x80	10	691.39	8821	2	730.58	2163	3	880.17	296.30	750.05	422.52
WRN-A-96x8x80	20	1548.05	24855	8	1613.55	551	2	1864.00	667.76	1669.25	1010.08
WRN-A-96x8x80	30	2435.64	10049	4	2513.95	318	3	2833.99	1078.13	2530.08	1740.70
WRN-A-96x8x80	40	3339.65	300	0	3414.16	54	1	3711.39	1522.24	3428.77	2599.22
WRN-B-96x12x40	10	669.73	16155	3	682.28	895	2	887.27	201.66	709.95	389.37
WRN-B-96x12x40	20	1532.12	3918	1	1558.96	259	2	1894.40	472.27	1545.83	1112.21
WRN-B-96x12x40	30	2435.38	4197	2	2443.20	1374	19	2963.43	788.62	2417.24	2144.23
Average		**1807.42**			**1850.95**			**2147.81**	**718.14**	**1864.45**	**1345.48**

Tabla 5: Resultados globales de los algoritmos de selección/agrupamiento.

	qv1	**qv2**	**qv3**	**qv4**	**sb-red**	**am-pol**	**cvs**
ARI	0.541	0.568	0.349	0.517	0.386	0.480	0.463
precision	0.607	0.523	0.557	0.498	0.705	0.551	0.573
recall	0.596	0.737	0.627	0.774	0.228	0.868	0.280

Tabla 6: Promedio de ARI en diferentes distribuciones contaminadas

	qv1	qv2	qv3	qv4	sb-red	am-pol	cvs
N(0,1)	0.730	0.774	0.586	0.651	0.481	0.630	0.783
Ncor0.5	0.055	0.330	0.190	0.321	0.454	0.328	0.783
Unif01	0.689	0.705	0.175	0.768	0.166	0.585	0.346
Gamma11	0.688	0.462	0.447	0.328	0.442	0.374	0.008

Tabla 7: Promedio de Precisión y Recall en diferentes distribuciones contaminantes.

	qv1	qv2	qv3	qv4	sb-red	am-pol	cvs
N(0,1)-Prec	0.799	0.619	0.663	0.515	0.832	0.628	1.000
Recall	0.752	0.859	0.685	0.831	0.233	0.921	0.477
Ncor0.5-Prec	0.072	0.304	0.302	0.367	0.797	0.419	1.000
Recall	0.092	0.490	0.347	0.608	0.218	0.799	0.477
Unif01-Prec	0.791	0.621	0.814	0.710	0.419	0.649	0.371
Recall	0.760	0.803	0.906	0.960	0.240	0.889	0.203
Gamma11-Prec	0.767	0.547	0.449	0.400	0.771	0.508	0.008
Recall	0.782	0.795	0.572	0.698	0.220	0.865	0.006

Tabla 8: Promedio de Precisión y Recall en diferentes distribuciones contaminantes.

	qv1	qv2	qv3	qv4	sb-red	am-pol	cvs
High	0.735	0.835	0.499	0.731	0.702	0.705	0.719
Medium	0.535	0.521	0.281	0.483	0.253	0.420	0.397
Low	0.352	0.348	0.269	0.338	0.202	0.314	0.233

Tabla 9: Precisión y recuperación medias de los grados de separación entre clusters.

	qv1	qv2	qv3	qv4	sb-red	am-pol	cvs
HighSep-Prec	0.645	0.584	0.579	0.516	0.903	0.642	0.742
Recall	0.615	0.812	0.645	0.814	0.322	0.888	0.413
MediumSep-Prec	0.619	0.508	0.541	0.480	0.665	0.525	0.551
Recall	0.573	0.727	0.598	0.780	0.195	0.880	0.246
LowSep-Prec	0.559	0.476	0.552	0.498	0.546	0.486	0.394
Recall	0.601	0.672	0.639	0.730	0.166	0.838	0.162

Tabla 10: Promedio de ARI en la relación de variables relevantes y enmascaradas.

	qv1	qv2	qv3	qv4	sb-red	am-pol	cvs
#mask = #true	0.571	0.696	0.443	0.669	0.427	0.663	0.486
#mask = 2(#true)	0.510	0.439	0.256	0.365	0.345	0.296	0.438

Tabla 11: Promedio de ARI en la relación de variables relevantes y enmascaradas.

	qv1	qv2	qv3	qv4	sb-red	am-pol	cvs
#mask = #true-prec	0.736	0.669	0.665	0.615	0.783	0.664	0.580
recall	0.531	0.859	0.681	0.928	0.238	0.842	0.286
#mask = 2(#true)-prec	0.479	0.377	0.449	0.381	0.627	0.439	0.565
recall	0.662	0.615	0.574	0.620	0.217	0.895	0.274

Tabla 12: Precisión y recuperación medias en relación con las variables relevantes y de enmascaramiento.