Trabajo de Clustering, Desarrollo de Sistemas Inteligentes

JAVIER CÓRDOBA ROMERO and JUAN JOSÉ CORROTO MARTÍN

En este trabajo se trata un problema tan antiguo como interesante, el problema de agrupar una serie de datos en distintos grupos, en este caso los datos pertenecen a los clientes de un distribuidor de mayorista de comida en el área de Portugal. Para clusterizar los datos se ha utilizado la técnica `DBSCAN` y los métodos aceptados en la literatura para intentar obtener la parametrización óptima, el resultado obtenido ha sido un sólo grupo donde se encuentran todos los datos. Como conclusión, la técnica `DBSCAN` no es útil para clusterizar este dataset.

Con el objetivo de extraer información de los datos, se han agrupado mediante loccalización y tipo de comercio, y se ha tratado de extraer aquellas características más interesantes, así como la importancia de cada grupo para el negocio.

Additional Key Words and Phrases: datasets, clustering, machine learning

ACM Reference Format:
Javier Córdoba Romero and Juan José Corroto Martín. 2020. Trabajo de Clustering, Desarrollo de Sistemas Inteligentes. 1, 1 (March 2020), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCCIÓN

2 HITO 1

2.1 Análisis de variables

El dataset con el que se trabajará contiene los datos de las ventas anuales de un distribuidor mayorista a 440 clientes, el objetivo planteado es el de determinar los patrones de clientes.

El dataset está compuesto por 8 variables, de las cuales 2 de ellas son cualitativas (Región del cliente y canal de distribución) mientras que las otras 6 son cuantitativas (ventas en número de euros de diferentes tipos de productos).

Como el objetivo perseguido es la determinación de patrones de clientes, y no de su ubicación o canal de distribución sólo se escogerán las variables cuantitativas, las que proporcionan datos sobre los tipos de productos vendidos.

2.2 Determinación de outliers

Antes de determinar los outliers del dataset fueron normalizados usando el *StandardScaler* de *sklearn*, la razón por la cual se realizó esta normalización fue con el objetivo de poder tener mayor granularidad a la hora de escoger el parámetro *epsilon* de *DBSCAN* cuando se usase el método de dibujar las distancias al k-ésimo vecino.

Para determinar los outliers del dataset se ha utilizado el algoritmo *Isolation Forest*, la ventaja de este algoritmo frente a otros, como por ejemplo *Local Outlier Factor* es que nos permite descubrir los outliers de un dataset sin la necesidad de ajustar los hiperparámetros de este, lo que agiliza el desarrollo de nuevos modelos.

Después de aplicar este algoritmo al dataset se identificaron 44 outliers, en la Figura 1 se pueden observar los outliers identificados una vez aplicado PCA al dataset. Sin embargo, esta Figura no es representativa del dataset ya que el ratio de varianza más alto es de 44%, cuando en clase se explicó que el mínimo de ratio de varianza a alcanzar es del 75%.

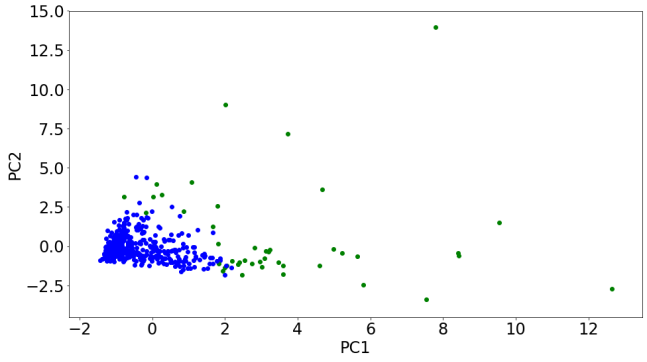

Fig. 1. Gráfico de dispersión aplicado a los datos de las 6 columnas después de aplicar PCA. Los puntos verdes son los identificados como outliers.

Una vez determinados los outliers, se procedió a comprobar si estos conformaban un grupo en sí mismos o si eran diferentes entre ellos mismos.

El resultado de esta comparación fué la de considerar a los outliers como un único grupo, ya que todos ellos tenían la misma característica: la mayoría de columnas con valores muy bajos y una o dos columnas con valores muy altos.

2.3 Principio de Pareto

Una vez los outliers fueron identificados y eliminados del dataset de trabajo se pasó a comprobar si se cumplía la regla de Pareto, esta regla enuncia que el 80% de ingresos del cliente mayorista proviene del 20% de clientes minoristas.

En este caso, esta regla no se cumple ya que el 20% de clientes suponen un 36% de los ingresos, como se puede ver en la Figura 2

2.4 Clustering inicial

DBSCAN se encuentra gobernado por dos parámetros, *min_pts* y *epsilon*, el objetivo de estos parámetros es controlar cuándo un punto es considerado parte del *core* de un clúster. Dicho de otro modo, un punto es considerado parte del *core* de un clúster cuando tiene `min_pts` puntos a una distancia menor de `epsilon` cerca de él

En la literatura se han descrito muchos métodos para obtener el valor óptimo de estos dos parámetros, en este estudio se utilizará el método de la "rodilla" para obtener el valor de *epsilon* [Sawant

Authors' address: Javier Córdoba Romero, javier.cordoba1@alu.uclm.es; Juan José Corroto Martín, JuanJose.Corroto@alu.uclm.es.

XXXX-XXXX/2020/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

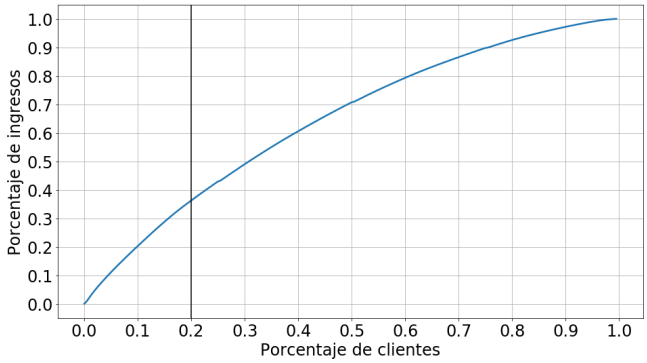

Fig. 2. Gráfico de línea donde se muestra el porcentaje de clientes (eje X) frente al porcentaje de ingresos (eje Y). La línea vertical negra muestra dónde tendría que tomar la gráfica el valor de 0.8 para que se cumpliese la regla de Pareto, en este caso, esta condición no se alcanza hasta el 60% de clientes.

2014] y el método de 2 * n_dimensiones para obtener el valor de *min_pts*[Schubert et al. 2017]

El primer método consiste en dibujar las distancias de todos los puntos a su k-ésimo vecino ordenados de forma ascendente y encontrar el punto donde la gráfica empieza a crecer muy rápidamente, es decir, donde la primera derivada es más alta, en la Figura 3 se puede ver un ejemplo de este método.

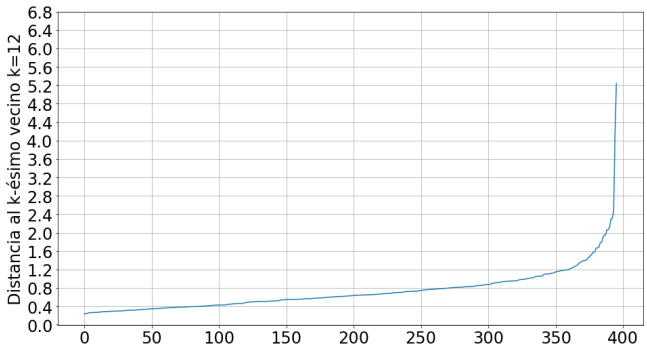

Fig. 3. Gráfico donde se muestra la distancia euclídea el k-ésimo vecino (k=12), como se puede ver, alrededor de 1.7 unidades de medida la gráfica pasa de crecer muy poco a crecer mucho, este es el valor óptimo de *epsilon* a tomar.

El método para escoger el valor de *min_pts* es más sencillo, la literatura recomienda que el valor sea `2 * nº dimensiones` donde el dataset de trabajo contiene 6 dimensiones por lo que el valor óptimo sería 12.

Una vez se saben los parámetros óptimos del algoritmo *DBSCAN* se pasa a realizar el clustering con un resultados muy pobres, el algoritmo sólo identifica un cluster de 393 datos y otros 3 datos que son considerados ruido y que nosotros consideramos como parte de los outliers que previamente no fueron identificados.

2.5 Optimización de hiperparámetros

Aunque el intento de clustering anterior, con los parámetros óptimos según la literatura no fuera exitoso se intentarán optimizar los hiperparámetros del algoritmo junto con la distancia usada.

En esta optimización se probarán valores de *min_pts* desde 2 hasta 30 con paso 2 y el valor de *epsilon* según el método de la rodilla. Este procedimiento se realizará tanto para la distancia euclídea como la de Chebyshev. La métrica a maximizar es la *silhouette*, una medida que tiene en cuenta la distancia intra-cluster.

Desafortunadamente, ninguna combinación de distancia, *min_pts* y *epsilon* consiguieron clusterizar en más de un grupo.

La conclusión extraída de todo este proceso es que *DBSCAN* no es algoritmo adecuado para este dataset, es decir, que en realidad existen varios grupos pero *DBSCAN* no es capaz de diferenciarlos.

3 HITO 2

Para el hito 2, hemos decidido no usar los resultados del clustering del hito anterior, puesto que no los consideramos representativos ni útiles. Con el fin de sacar la máxima información posible de los datos, hemos decidido generar grupos "a priori" con la información de las variables categóricas que descartamos anteriormente. Los datos que se han considerado para este hito son los resultantes de eliminar los outliers, tal y como se ha explicado anteriormente.

De esta forma, se han generado 6 grupos, atendiendo a cada combinación posible de las variables *Channel* y *Region*. Los grupos han sido codificados de la siguiente manera:

0 *Channel* 1 y *Region* 1. Distribuidor minorista en la zona de Lisboa.
1 *Channel* 2 y *Region* 1. Hotel, cafeterías, etc. en la zona de Lisboa.
2 *Channel* 1 y *Region* 1. Distribuidor minorista en la zona de Oporto.
3 *Channel* 2 y *Region* 2. Hotel, cafeterías, etc. en la zona de Oporto.
4 *Channel* 1 y *Region* 2. Distribuidor minorista en otras zonas.
5 *Channel* 2 y *Region* 2. Hotel, cafeterías, etc. en otras zonas.

A modo de explorar de nuevo los datos con los grupos recién descritos, se ha realizado un análisis de componentes principales. El resultado se muestra en la figura 4. A parte de tener en cuenta que los ejes no son muy representativos, debido a la baja varianza que el algoritmo reporta, también se puede ver que los datos se encuentran probablemente mezclados.

3.1 Cálculo de representantes de grupos

Para el cálculo de los representantes, hemos decidido calcular la *mediana* de cada columna, debido a que es una medida estadística menos sensible a datos extremos. Esta decisión viene fomentada debido a la alta variación de cada grupo, que como se puede ver ligeramente en 4, cada grupo se encuentra disperso y ningún grupo se diferencia especialmente del resto.

3.2 Descripción semántica de grupos

En esta sección se ofrece una descripción semántica de cada grupo, más allá de categorizarlos según las variables cualitativas. Para ello, se han realizado tests de *Kruskal-Wallis* para cada una de las

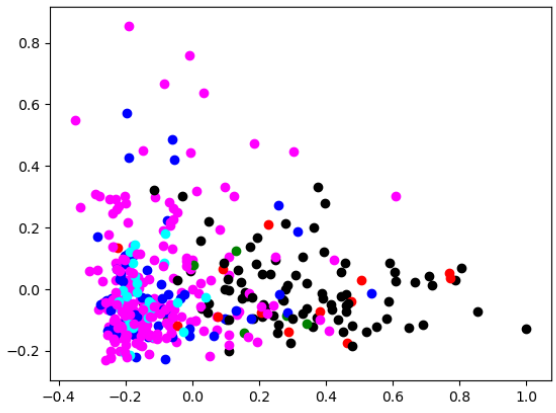

Fig. 4. Gráfica PCA con los datos coloreados dependiendo de su grupo.

Group	Delic	Det_Pap	Fresh	Froz	Groc	Milk
0	710.0	412.0	8656.0	1801.0	2501.0	2204.0
1	1414.0	3441.0	2665.5	2352.5	7650.5	5955.0
2	868.0	299.0	9784.0	2679	3315.0	1511.0
3	806.0	4111.0	6468.0	902.0	9965.0	6243.0
4	804.5	372.5	9581.5	1870.5	2061.0	2232.0
5	1282.0	4508.0	7362.0	993.0	1109.0	7027.0

Table 1. Representantes de cada grupo.

variables, y se han utilizado los box plots resultantes del test para intentar describir de forma semántica cada grupo. A continuación, un breve comentario de cada una de las variables:

Delicassen. El resultado del test de *Kruskall-Wallis* sobre esta variable muestra un p-valor de 0.4335. Por tanto, esta variable no muestra diferencias en ninguno de los grupos (Como se puede ver también en la figura 5). Todos los grupos parecen gastar más o menos el mismo dinero en productos delicatessen.

Detergents_Paper. Para esta variable, el test muestra un p-valor de $1.34E-38$. También se puede ver en la figura 6 cómo hay diferencias significativas respecto a las distribuciones de la variable. Es más, llama la atención la similaridad de los grupos pares (aquellos que representan distribuidores minoristas en las 3 regiones) respecto a los impares, representando Hoteles y cafeterías. El resultado de este test nos muestra que, efectivamente, los hoteles y cafeterías hacen un gasto mucho mayor en productos de limpieza que distribuidores minoristas, siendo los hoteles y cafeterías de Lisboa (grupo 1) los que menos gasto hacen.

Fresh. Esta es la única variable aparte de *Delicassen* que presenta un p-valor mayor de 0.05, en este caso 0.073. Aunque en el box plot (Figura 7) se muestran ligeras diferencias en distribución entre grupos, no son suficientes como para ser significativas.

Frozen. Para esta variable, el test reporta un p-valor de $4.29E-5$. En la figura 8 se puede ver como los grupos pares vuelven a exhibir comportamientos similares, mientras que los grupos representantes de hoteles y cafeterías se distancian un poco, esta vez consumiendo menos. Esto podría significar que los hoteles y cafeterías compran pocos productos congelados, siendo esto una excepción para los de la zona de Lisboa.

Grocery. Esta variable muestra distribuciones similares a la de productos de limpieza. Parece ser que los hoteles y cafeterías consumen significativamente más productos de alimentación genéricos, siendo los de Lisboa los que menos y aquellos pertencientas a otras regiones (es decir, diferentes de Lisboa y Oporto) los que más. El p-valor para este test fue $2.31E-35$

Milk. En esta variable también se muestran las diferencias mostradas en las demás. Los hoteles y cafeterías de todas las regiones muestran consumos completamente diferentes a los de distribuidores minoristas. En este caso, son los hoteles y cafeterías de oporto los que más suelen consumir. El p-valor de este test fue $1.19E-25$.

Llama mucho la atención cómo en casi todas las variables existe una gran similaridad entre grupos pares e impares. Esto podría suponer que la zona geográfica no es un buen diferenciador, siendo el más importante el tipo de cliente que sea.

Usando los resultados de los tests anteriores se obtienen las siguientes conclusiones sobre cada grupo:

Grupo 0. Este grupo representa distribuidores minoristas de la zona de Lisboa. Llama la atención la similaridad que tiene respecto a otros distribuidores minoristas, teniendo prácticamente la misma distribución en cada variable. La única variable que parece significativamente diferente es *Frozen*, para la que parece mostrar un mayor gasto en general que el resto de distribuidores.

Grupo 1. Los hoteles y cafeterías de Lisboa presentan comportamientos similares a otros hoteles y cafeterías. Sin embargo, en algunas variables en las que los grupos impares mostraban mayor consumo, los de Lisboa eran los que menos mediana presentan. Ejemplos en Figuras 9, 6. Sin embargo, son los únicos hoteles y cafeterías que no muestran una reducción importante de productos congelados.

Grupo 2. Los distribuidores minoristas de Oporto muestran un comportamiento similar a los de otras regiones. Por lo general, son los que menos gastos hacen en casi todos los tipos de productos, con excepción en productos frescos.

Grupo 3. De nuevo, los hoteles y cafeterías de Oporto tienen un comportamiento similar al resto de hoteles. Llama la atención la gran variación que presentan en productos derivados de la leche, siendo ellos los que más consumen generalmente. También llama la atención el poco gasto que hacen en productos congelados.

Grupo 4. La mayor característica del grupo 4 no son sus gastos, que son prácticamente similares al resto de grupos pares. Lo que más llama la atención de este grupo son la cantidad de datos atípicos que presentan en todas las variables. Se podría pensar que existen clientes específicos que gastan más que el resto de distribuidores minoristas en zonas diferentes a Lisboa y Oporto.

Grupo 5. De nuevo, un grupo impar con comportamientos similares a otros grupos impares. Llama la atención cómo este grupo presenta, por lo general el mayor gasto en aquellas variables en las que los hoteles y cafeterías se desmarcan. Suelen presentar la mediana más alta, y por tanto, son los que más dinero suelen gastar.

3.3 Determinación de importancia de cada grupo

Para determinar la importancia de cada grupo, hemos utilizado la información de las secciones anteriores, así como un cálculo del gasto total y medio de cada uno. Estos gastos se pueden ver en la tabla **??**. En ella se puede ver cómo los grupos impares tienen bastante más gasto medio, aunque alguno de ellos tenga menos gasto total, debido al número de datos en cada grupo.

Group	Gasto total	Número de clientes	Gasto medio
0	1422203.0	57	24950.93
1	170743.0	6	28457.17
2	588327.0	27	21789.89
3	442648.0	13	34049.85
4	5302440.0	208	25492.5
5	3076577.0	85	36195.02

Table 2. Gasto total y medio de cada grupo.

Cómo resultado del hito 2, y para explicar también qué grupos son más importantes, se presentan las siguientes conclusiones:

- Los hoteles y cafeterías son los que más gasto medio presentan. De entre ellos destaca el grupo 5 (hoteles y cafeterías de sitios diferentes a Lisboa y Oporto). Este grupo es el que consideramos más importante porque es el que más beneficios da.
- Destacan la cantidad de clientes que son distribuidores minoristas, sobre todo el grupo 4. Si tuvieramos que identificar el grupo más importante de estos distribuidores minoristas, sería el 4.
- A pesar de tener muchos más datos que los grupos 1 y 3, el grupo 2 tiene casi los mismos gastos totales. Esto es debido a que tiene el gasto medio menor de todos. Podría considerarse este grupo como el menos importante.

REFERENCES

Kedar Sawant. 2014. Adaptive Methods for Determining DBSCAN Parameters.
Erich Schubert, Jörg Sander, Martin Ester, Hans Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems* 42 (07 2017), 1–21. https://doi.org/10.1145/3068335

A GRÁFICAS COMPLEMENTARIAS

En este anexo se muestran gráficas resultado de aplicar algunos de los experimentos explicados en el resto del artículo.

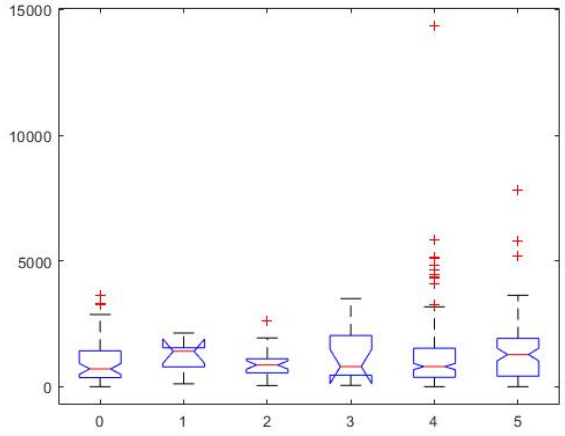

Fig. 5. Box plot resultado del test no paramétrico sobre la variable *Delicassen*.

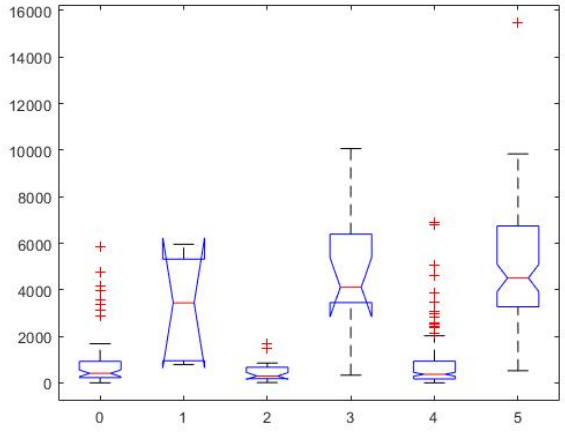

Fig. 6. Box plot resultado del test no paramétrico sobre la variable *Detergetents_Paper*.

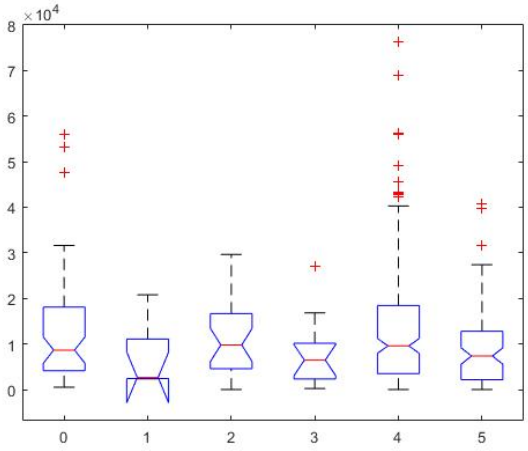

Fig. 7. Box plot resultado del test no paramétrico sobre la variable *Fresh*.

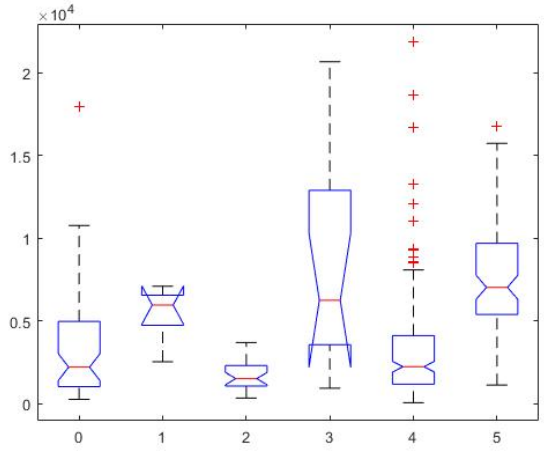

Fig. 10. Box plot resultado del test no paramétrico sobre la variable *Milk*.

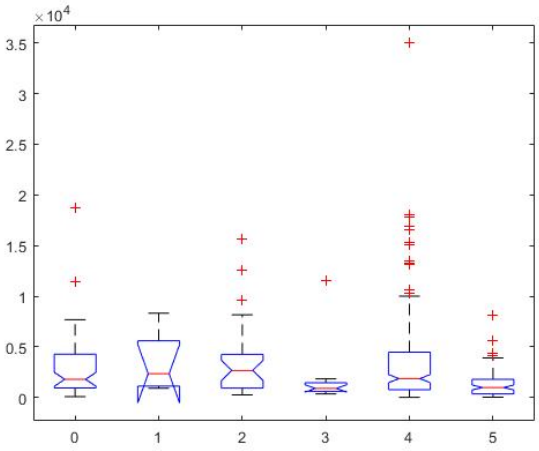

Fig. 8. Box plot resultado del test no paramétrico sobre la variable *Frozen*.

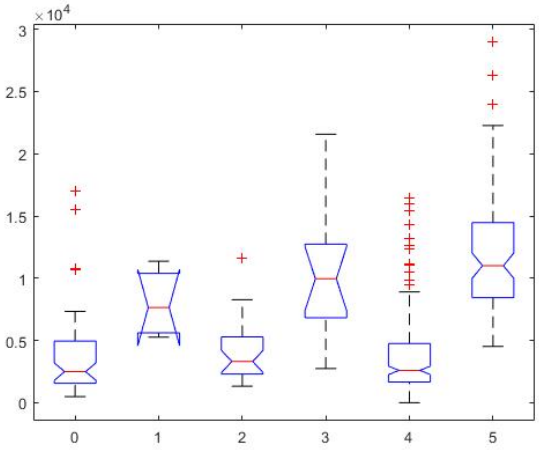

Fig. 9. Box plot resultado del test no paramétrico sobre la variable *Grocery*.