Unsupervised learning using Self-organizing maps for client's groups

FELIX ANGEL MARTINEZ MUELA* and MIGUEL DE LA CAL BRAVO*, Universidad de Castilla La Mancha, Spain

Proyecto para la asignatura del Máster en Ingeniería Informática llamada Desarrollo de Sistemas Inteligentes, en la cual, se emplearán R y Matlab para realizar un análisis de los datos de ventas de un mayorista para determinar los grupos de clientes que tiene empleando la técnica Mapas Auto-organizados.

CCS Concepts: • **Computing methodologies → Cluster analysis**;

Additional Key Words and Phrases: Self-organizing maps, unsupervised learning

ACM Reference Format:
Felix Angel Martinez Muela and Miguel de la Cal Bravo. 2020. Unsupervised learning using Self-organizing maps for client's groups. *ACM Trans. Graph.* 0, 0, Article 0 (March 2020), 4 pages.

1 INTRODUCCIÓN

En este trabajo se realiza un análisis cluster sobre un dataset que contiene los gastos por clientes según diferentes tipos de productos. Este dataset está formado por un total de 440 filas, que representan cada uno de los clientes, así como de 8 columnas o atributos de cada uno de los mismos.

2 HITO 1: DETERMINACIÓN DE CLUSTERS

En este primer hito, tenemos por objetivo principal descubrir patrones de clientes en un conjunto de datos dado. Para hacer esto posible, realizaremos un análisis cluster que nos permitirá obtener un conjunto de grupos de clientes con características similares dentro de los elementos que conforman cada grupo.

2.1 Análisis de variables

Conocidas las variables de nuestro dataset, hemos decido quedarnos con aquellas variables cuantitativas (columnas de la C a la H), descartando las variables cualitativas (región y tipo cliente), ya que éstas nos resultan indiferentes a la hora de sacar conclusiones y determinar semánticas de clusters.

De esta manera, únicamente trabajaremos con los gastos anuales de cada cliente en cada uno de los diferentes productos (frescos, lácteos, ultramarinos, congelados, detergentes y papelería, y en productos delicatessen) o en la suma acumulada de todos ellos.

*Both authors contributed equally to this research.

Authors' address: Felix Angel Martinez Muela, FelixAngel.Martinez@alu.uclm.es; Miguel de la Cal Bravo, Miguella.Cal@alu.uclm.es, Universidad de Castilla La Mancha, Ciudad Real, Ciudad Real, Spain, 13005.

0730-0301/2020/3-ART0 $15.00
https://doi.org/

2.2 Determinación de Outliers

En este apartado hablamos de aquellos clientes anómalos, con valores muy dispersos en las variables (muy buenos en general, o en un tipo de producto determinado).

De cierta manera, al aplicar el algoritmo que nos ha sido asignado, hemos podido detectar que se forma un pequeño grupo que entendemos que serían los outliers (se corresponden con los clientes con mayor número de ventas).

El problema aquí, es estudiar si estos elementos forman parte de un grupo o si simplemente son outliers que han sido agrupados para diferenciarlos del cluster principal.

De todas maneras, nosotros optamos por no tener en cuenta estos clientes para el análisis cluster, eliminando así el "ruido" que estos pudieran generar sobre el resto de los datos.

Para la determinación de *outliers* hemos empleado una técnica de detección mediante **Jacknife**.

Se han detectado un total de 7 outliers. Las empresas en las filas: **86, 87, 182, 285, 326, 334 y 406**, se consideran outliers debido a la diferencia de sus valores respecto al resto de empresas.

2.3 Core del negocio. Principio de Pareto

Para determinar el núcleo de negocio, realizaremos sumas de los gastos totales en productos de cada uno de los clientes, para posteriormente ordenarlos y quedarnos con el 20% de los clientes que más ventas han generado.

Tras realizar esta suma total, nuestra misión consiste en comprobar si ese 20% de clientes genera el 80% del total de los costes del negocio. Gracias a esta regla, lograríamos identificar un % muy pequeño de nuestros clientes que supone un gran % de los costes globales.

Aplicando el algoritmo hemos determinado que **NO** se cumple el principio de Pareto, ya que con el 20% (88 primeros) de los clientes que más compran obtenemos aproximadamente el **43%** de las ventas, y luego vimos que tendríamos que coger el 43% (251 primeros) del total de clientes para que eso se cumpliese.

De esta manera, concluimos afirmando que ni mucho menos se cumple la regla de 20-80 o principio de Pareto.

2.4 Realización del análisis cluster y número de grupos óptimo

En este apartado haremos uso del algoritmo de clustering que nos ha sido asignados, en este caso, dicho algoritmo es la de **Mapas Autoorganizados (ó *Self-Organized Maps, SOM*)**. A la hora de utilizar este algoritmo, debemos considerar las siguientes variables para crear el grid:

- `xdim`: número de filas de celtas.
- `ydim`: número de columnas de celdas.
- `topo`: topología del grid, puede ser hexagonal, circular, etc.

A continuación, en la figura 1, se muestra un gráfico obtenido con R, gracias a la librería de *Kohonen*:

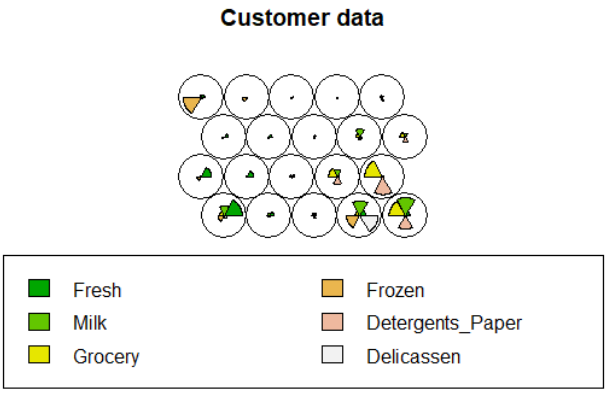

Fig. 1. Resultado de aplicar Mapas Autoorganizados, en una malla de 5x4.

Mediante este algoritmo obtenemos una manera muy peculiar de agrupar los datos, en base a sus características. También resulta interesante saber cuántos elementos pertenecen a cada grupo, lo cual se ve representado en la figura 2:

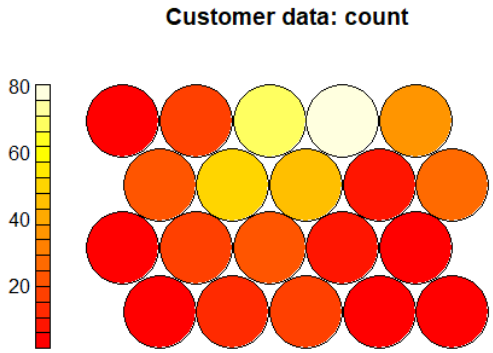

Fig. 2. Número de elementos en cada celda de la malla.

De esta manera podremos ver la correspondencia entre los grupos obtenidos, y el número de elementos que pertenecen a cada uno de ellos. Llegados a este punto, resulta interesante determinar la parametrización óptima, en este caso basado principalmente en el número de filas y columnas que compondrán nuestra malla.

En cuanto al número óptimo de grupos a efectuar, hemos optado por utilizar un **criterio de información bayesiana BIC**. Tras la aplicación del algoritmo, obtuvimos que el número óptimo de grupos con k-means era de 6 o 7, aunque nosotros posteriormente decidimos quedarnos con únicamente 2 grupos ya que la diferencia de similitud era muy pequeña comparando con los 6-7 grupos que se formaban.

2.5 Parametrización óptima del algoritmo

Para concluir el hito 1, se determina la parametrización óptima del algoritmo de Mapas Autoorganizados para la formación de los clusters.

En este caso, hemos desarrollado un algoritmo que hace uso de la técnica de **Silhouette**, combinando diferentes tamaños del grid (xdim, ydim, vistos en el apartado anterior), así como la forma de dicho grid. El resultado de dicho algoritmo se muestra en la figura 3.

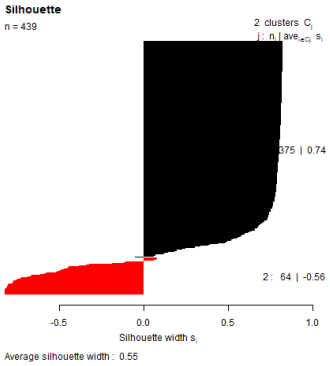

Fig. 3. Resultado de aplicar la técnica de Silhouette.

Al tratarse de un dataset muy particular, que no tiene grupos claramente diferenciados, con dicha técnica no hemos obtenido buenos resultados para aplicarlos a SOM. Con un grid pequeño el coeficiente de Silhouette resultante es muy bueno, pero el propio algoritmo SOM no termina convergiendo.

Finalmente, obtuvimos que la mejor parametrización se obtiene con un `xdim=1` e `ydim=2` con forma **hexagonal**, teniendo un total de 2 celdas pero sin llegar a converger el algoritmo.

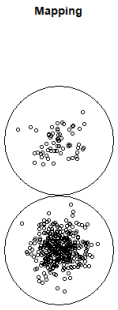

Fig. 4. Mapas Autoorganizados. Representación gráfica de la formación de dos grupos, con una fila y dos columnas de celdas y forma hexagonal.

Así, concluimos el primer hito donde hemos partido del dataset inicial, analizamos las variables que participan en el estudio, determinamos los outliers, vimos que no se cumplía el principio de Pareto y, para finalizar, realizamos un análisis cluster con el algoritmo de Mapas Autoorganizados estudiando su parametrización óptima.

3 HITO 2: SEMÁNTICA DE LOS CLUSTERS

En el segundo hito, nuestro objetivo es comprender correctamente los grupos que se han determinado en el anterior hito. Para ello, trataremos de caracterizar a dichos cluster con sus propiedades esenciales.

3.1 Cálculo de los representantes de los grupos

Además de calcular los representantes de los grupos, también se persigue el objetivo de caracterizar los mismos, hallando las características que los definen.

En este caso, obtenemos un grupo principal con la mayoría de clientes, los cuales tienen una media de ventas dentro de lo normal.

Por otro lado, tenemos otro grupo diferenciado que se encuentra formado por aquellos clientes con un nº de ventas fuera de lo normal, ya que son aquellos que acumulan el mayor número de costes anuales o aquellos con ventas mínimas.

En cuanto a los representantes de los grupos, hallamos la media de ventas para determinar los mismos. En la tabla 1, podemos ver los representantes de los clusters.

Table 1. Representantes de los clusters

Grupo	Media ventas	Descripción
Grupo principal	**11065.76**	Mayor media de ventas
Grupo secundario	4589.24	Menor media

Como vemos en la tabla, un representante del grupo principal sería un cliente con una media de 11065 por producto, mientras que el segundo grupo cuenta con una media de 4589.

Esto significa que, el primer grupo, de media, vende 2.41 veces más que el segundo, algo muy representativo en este cluster.

3.2 Importancia de cada grupo para la empresa

En este momento, nos interesa destacar la importancia de cada uno de los grupos obtenidos, para ver así que nos aportan como empresa. Esto podríamos hacerlo de dos formas:

- Gastos totales por **grupos de clientes**.
- Gastos totales por **productos**.

Vistos los resultados obtenidos en el hito 1, en este apartado podemos decir que tenemos un cluster principal con clientes que venden todo tipo de productos, pero no podemos ser nada más concretos con un dataset tan "especial".

Una curiosidad que hemos extraído al aplicar nuestro algoritmo es que, de los dos grupos que se forman, la media en cinco de los seis productos más vendidos se encuentran en el primer grupo, mientras que en el secundario esta media es superior a la del primero en productos de detergentes y papeles.

En la figura 5, podemos observar la agrupación obtenida al emplear el algoritmo de mapas autoorganizados, algo que se corresponde con la media extraída (ver tabla 2) de los productos en ambos grupos.

Viendo la media de ventas entre los dos grupos, observamos que en ambos el tipo de producto más vendido son los frescos, con una amplia diferencia en el grupo primero con respecto al resto.

Table 2. Media de ventas por tipo de producto por grupos

Tipo producto	Media ventas grupo 1	Media ventas grupo 2
Frescos	**36592.917**	8096.145
Leche	**9084.717**	5169.069
Comestibles	**8615.183**	7622.351
Congelados	**6679.683**	2506.182
Det.Papel	2177.133	**2892.881**
Delicatessen	**3244.900**	1248.826

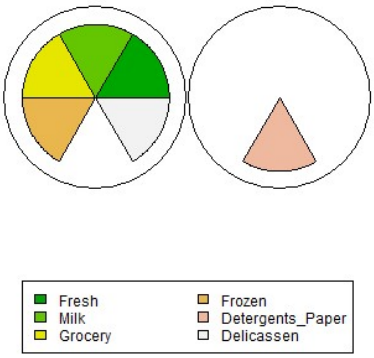

Fig. 5. Mapas Autoorganizados. Representación gráfica por ventas de los diferentes productos.

Además, este tipo de producto es el que más veces se vende comparando ambos grupos, pues se vende 4.52 veces más en el primer grupo con respecto al segundo.

Para terminar, también se detecta en dicho grupo que los productos frescos acumulan más ventas de media que la suma de las medias del resto de tipos de productos del grupo, así como también es superior a la suma de la media de ventas de todos los productos del segundo grupo, algo de más importancia si cabe tras haber quitado los outliers.

3.3 Estudio estadístico de variables significativamente diferentes

Siguiendo en la línea del punto anterior, en el dataset estudiado no se ha podido encontrar variables significativamente diferentes que conformen los clústers.

En este apartado aplicamos algunas de las herramientas estadísticas y tests no paramétricos.

3.3.1 Contraste de hipótesis. Debido a que entre nuestros datos tenemos la característica de la zona en la que dicho cliente se encuentra, hemos decidido comprobar si tiene alguna relación el PIB de dicha ciudad respecto a al producto que se compra, por lo que hemos planteado las siguientes hipótesis:

- `Hipótesis nula`: la media de los productos delicatessen que se compran en Lisboa es igual a la media global, ya que el PIB en dicha ciudad es el más alto de Portugal.

- `Hipótesis alternativa`: ser la ciudad con el PIB más alto no influye a la hora de tener el mayor gasto en productos delicatessen.

El *nivel de confianza* que vamos a tomar para que nuestra hipóteis nula sea rechazada es p-valor <= 5%. Después de aplicar p-valor a dichas hipótesis nos sale un p-valor de: **0.2711** por lo tanto se cumple nuestra hipótesis nula de que Lisboa al tener el PIB más alto de Portugal y por tanto mayor poder adquisitivo entre sus habitantes, compra más productos delicatessen.

3.4 Descripción semántica de los grupos obtenidos

Para concluir con el presente trabajo de clustering, procedemos a realizar una descripción semántica de los grupos que hemos obtenido al aplicar el clustering.

Grupo/Cluster principal: incluye la mayoría de clientes, con ventas generales en todos los productos sin poder distinguirse (o parecerse) lo suficiente para dar lugar a nuevos clusters. Lo que sí hemos comprobado es que su media de ventas es superior en todos los tipos de productos menos en uno en concreto.

Grupo/Cluster secundario: este segundo grupo vende de media menos que el primero, pero en cuanto a los productos de detergentes/papel vimos que sí vendía algo más este segundo grupo.

Hemos visto que este no es el mejor de los dataset para sacar conclusiones de los clientes, pues no se detecta una homogeneidad clara entre un gran número de clientes. Sin embargo, sí hemos sido capaces de descubrir aquellos clientes que más beneficios nos aportarían a la empresa de media, algo realmente importante para el negocio mediante un trato más particular frente al resto de clientes.

A EJECUCIÓN

Dado que se han empleado dos lenguajes que son: *R* y *Matlab*, los archivos se deben ejecutar en el orden adecuado, ya que de no hacerlo así, y por tanto no seguir el flujo correcto, los resultados serán distintos y erróneos, ya que la entrada de uno es la salida del paso previo. Quedando aclarado esto, el orden adecuado es:

(1) **1_pca.R**
(2) **2_Jackknife.m**
(3) **3_dsi_clustering_som.R**

En el archivo primero llamado **1_pca.R** se realiza un pequeño análisis de los datos para empezar a trabajar con ellos y posteriormente se realiza un análisis de componentes principales para de ésta manera, pasar de las 6 variables a una menor dimensión la cual nos permita una visualización de los datos en 2 dimensiones sin perder demasiada relación entre las variables representantes **pc1** y **pc2** y las reales. Una vez que tenemos dichas dos variables procedemos a exportarlas a un archivo excel llamado *pca_pc1pc2.csv*.

En el archivo segundo llamado **2_Jackknife.m** partiendo de lass dos variables principales anteriores para proceder a realizar un análisis de los outliers presentes y señalarlos. Para poder hacer esto, lo que hemos realizado ha sido probar K-means con un determinado número de clusters, y aplicar BIC (*bayesian information criterion*), lo cual nos indica el número óptimo de clústers en nuestro algorítmo. Una vez que hemos determinado el número de cluster que tendrá nuestro k-means que aplicaremos para sacar outliers, aplicamos **Jackknife** y esto definitivamente nos señala los outliers en nuestro algoritmo. Una vez que tenemos el número de empresa que se considera outlier, lo exportamos para proseguir con nuestro análisis.

En el último archivo llamado **3_dsi_clustering_som.R** eliminamos los outliers de nuestros datos, y procedemos a guardarlos en un archivo llamado *data_outliers.csv* para poder realizar un análisi posterior. Después aplicamos el Principio de Pareto. Después de esto aplicamos el *Coeficiente de Silhouette* y obtendremos de ésta manear nuestro mapa autoorganizado más coherente a grupos más parecido entre los datos internos y más distantes entre externos. Posteriormente realizamos las gráficas y análisis posteriores. Hemos realizado también un contraste de hipótesis en dicho script para poder sacar conclusiones respecto a los datos.