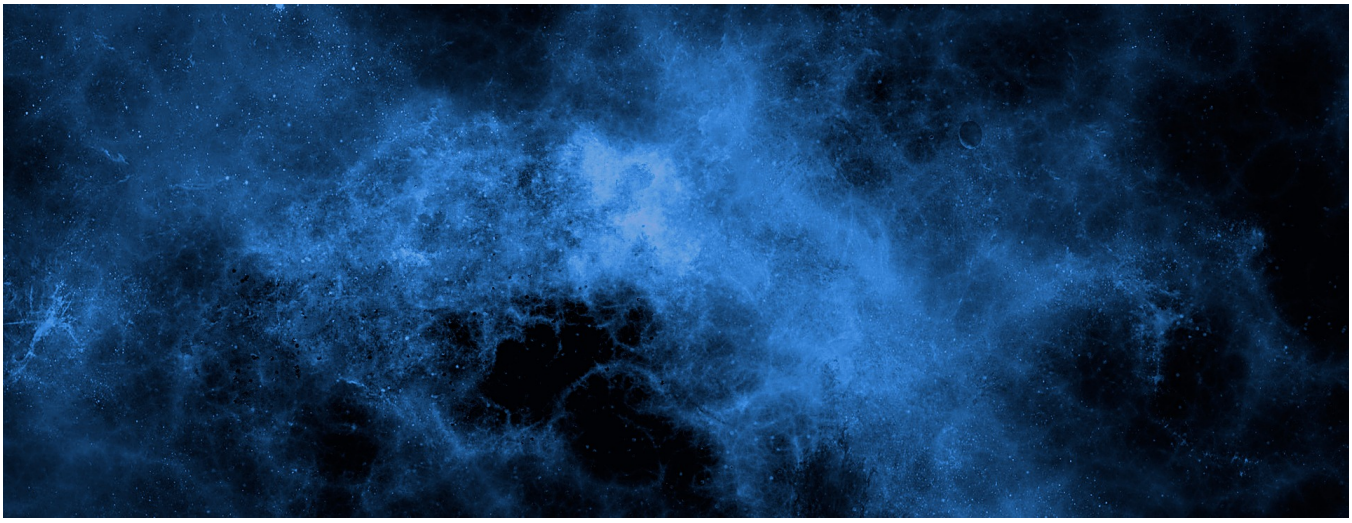
Trabajo de Clustering

Diego Andérica Richard
Diego.Anderica@alu.uclm.es
Escuela Superior de Informática
Ciudad Real, España



1 INTRODUCCIÓN

Este trabajo se centrará en el estudio de los datos proporcionados, que corresponden a una encuesta realizada en Holanda acerca de los transportes de los ciudadanos entre 2010 y 2012. Además, se debe realizar una tarea de *clustering* utilizando la técnica de *Fuzzy C-Means* asignada, donde cada uno de los elementos puede pertenecer a una serie de *clusters* en mayor o menor medida de acuerdo a una matriz de pertenencia.

El lenguaje en el que se ha desarrollado este trabajo ha sido Python, utilizando las diferentes librerías de las que dispone para este tipo de problemas, como Pandas o skfuzzy.

2 LECTURA Y VISUALIZACIÓN DE LOS DATOS

En primer lugar, se ha cargado el archivo CSV con la librería Pandas para, posteriormente, observar cómo se encuentran estructurados los datos (Figura 1). Además, se ha hecho un recuento de todas las características y elementos totales de los que se compone el *dataset*, obteniendo un total de **17 características** y **230.608 elementos**.

	mode_main	distance	density	age	male	ethnicity	education	income	cars	license	bicycles	weekend	diversity	green	temp	precip	wind
0	walk	1.0	1.26259	84	no	native	lower	less20	0	yes	1	yes	1.24604	26.881233	0.1	0.10	3.0
1	walk	10.0	1.26259	84	no	native	lower	less20	0	yes	1	yes	1.24604	26.881233	0.1	0.10	3.0
2	car	3.0	1.76264	27	yes	western	middle	20to40	1	yes	2	yes	1.53959	36.045955	-3.4	0.05	1.8
3	car	3.0	1.76264	27	yes	western	middle	20to40	1	yes	2	yes	1.53959	36.045955	-3.4	0.05	1.8
4	car	61.5	1.76264	27	yes	western	middle	20to40	1	yes	2	yes	1.53959	36.045955	-3.4	0.05	1.8

Figure 1: Estructura del Archivo

Como se puede observar en la Figura 1, algunas de las características son cualitativas, por lo que se ha procedido a la obtención de los tipos de cada una. Atendiendo a los resultados (Figura 2), se observa que las características **cualitativas** son: **mode_main**, **male**, **ethnicity**, **education**, **income**, **license** y **weekend**. Además, en la Figura 3 se pueden observar los diferentes valores que toman cada una de ellas, junto con el número total de veces que aparecen en el *dataset*.

```
mode_main      object
distance      float64
density       float64
age             int64
male           object
ethnicity      object
education      object
income         object
cars            int64
license        object
bicycles        int64
weekend        object
diversity     float64
green         float64
temp          float64
precip        float64
wind          float64
dtype: object
```

Figure 2: Tipos de Datos

```
Valores diferentes para la característica 'mode_main':
car     127439
bike     56298
walk     37571
pt        9300
Name: mode_main, dtype: int64


Valores diferentes para la característica 'male':
no     125676
yes    104932
Name: male, dtype: int64


Valores diferentes para la característica 'ethnicity':
native        201561
western        17772
nonwestern     11275
Name: ethnicity, dtype: int64


Valores diferentes para la característica 'education':
middle    88306
higher    79185
lower     63117
Name: education, dtype: int64


Valores diferentes para la característica 'income':
more40    106182
20to40     97140
less20     27286
Name: income, dtype: int64


Valores diferentes para la característica 'license':
yes    206986
no      23622
Name: license, dtype: int64


Valores diferentes para la característica 'weekend':
no     189250
yes     41358
Name: weekend, dtype: int64
```

Figure 3: Valores Características

Atendiendo a estos resultados, se deprende que la mayoría de viajes se realizaron con el coche como **medio de transporte**, mientras que el transporte público fue el menos utilizado para los viajes contenidos en es *dataset* proporcionado. Junto con esto, la **etnia** que más destaca es *native*. La cantidad de **hombres y mujeres** es similar, así como el **nivel de educación**. Por último, se registraron más viajes **entre semana**, hay muchas más personas que tienen el **carnet** de conducir que las que no y hay menos personas que tienen un nivel de ingresos bajo.

Por otra parte, se ha decidido obtener los valores estadísticos para cada una de las características cuantitativas, como se puede observar en la Figura 4. Estas estadísticas indican que, por ejemplo, la distancia media de los desplazamientos se sitúa en torno a 23.5 km, aunque hay al menos un trayecto de hasta 400 km. La edad de los participantes varía entre 18 y 98 años, la mayor parte de las personas poseen entre 1 y 2 vehículos y, en cuanto a la temperatura y precipitación, el rango de valores es relativamente amplio. Esto podría indicar que las encuestas se realizaron en diferentes estaciones del año entre 2010 y 2012.

	count	mean	std	min	25%	50%	75%	max
distance	230608.0	12.217913	23.545686	0.10000	1.50000	4.000000	12.000000	400.000000
density	230608.0	1.569055	1.593292	0.00184	0.59132	1.153210	1.952960	11.442960
age	230608.0	47.661356	15.934884	18.00000	36.00000	47.000000	60.000000	98.000000
cars	230608.0	1.382584	0.822056	0.00000	1.00000	1.000000	2.000000	10.000000
bicycles	230608.0	3.357134	1.936614	0.00000	2.00000	3.000000	4.000000	10.000000
diversity	230608.0	1.774927	0.493037	0.00000	1.38894	1.827390	2.172380	2.827560
green	230608.0	54.939470	22.172372	0.00000	37.11120	54.102123	74.381844	97.813002
temp	230608.0	13.316930	7.565732	-9.00000	8.00000	13.400000	19.000000	35.900000
precip	230608.0	2.184633	4.674801	0.00000	0.00000	0.100000	2.300000	142.300000
wind	230608.0	4.097702	1.914821	0.40000	2.70000	3.800000	5.100000	16.300000

Figure 4: Información Estadística Características Cuantitativas

3 TRATAMIENTO Y GRAFICADO DE LOS DATOS INICIALES

Una vez obtenidos los datos estadísticos de las diferentes características, se observó que algunas de ellas, como *distance*, *density* o *precip* alcanzan valores máximos relativamente alejados de la media. Por tanto, se decidió graficar los histogramas correspondientes a cada una de las características observado los resultados (Figura 5).

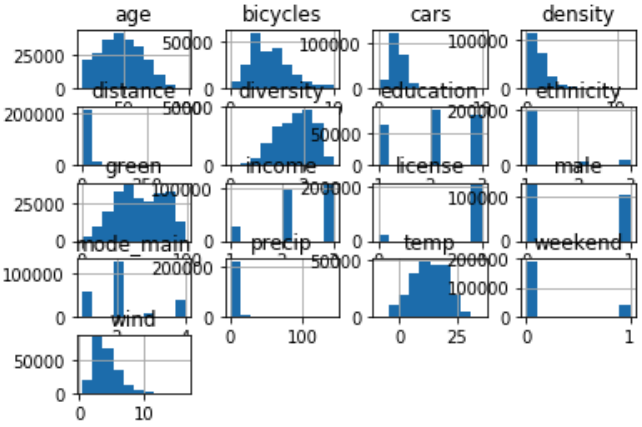


Figure 5: Histogramas Características

Por tanto, con el fin de eliminar la mayor cantidad de *outliers* posible, de manera que estos no influyeran en la etapa de *clustering*, se han descartado los elementos que se encontrasen más allá de 3 veces la desviación estándar con respecto a la media de los valores. Además, se han tratado y modificado las características cualitativas siguiendo una serie de reglas antes de la estandarización:

- Características que toman valores sí/no: se han sustituido estos por los números 1/0, respectivamente.
- Características que toman más de dos valores y son diferentes de sí/no: se han asignado números indistintamente a partir del 1 para las que no representan orden. Por otra parte, para las que siguen un cierto

orden, como puede ser la característica *income* (*less20*, *20to40*, *more40*), se han asignado valores crecientes a partir del 1 según la cantidad de ingresos (1 para *less20*, 2 para *20to40* y 3 para *more40*). De esta manera, se podría seguir manteniendo una relación numérica entre ellos sin perder el significado.

Finalmente, una vez finalizadas la eliminación de *outliers* y transformación de características cualitativas, se ha procedido a realizar una normalización o estandarización de los valores existentes, con el fin de que todos los valores posean la misma importancia (por ejemplo, *distance* vs *diversity*. Para ello, se ha utilizado la técnica de escalado MinMax, una de las más extendidas, haciendo que los valores oscilen entre 0 y 1.

Por tanto, con el *dataframe* preprocesado, cambiando características cualitativas por cuantitativas, realizando una eliminación de *outliers* y, finalmente, realizando un escalado, se podía proceder a realizar la tarea de *clustering*.

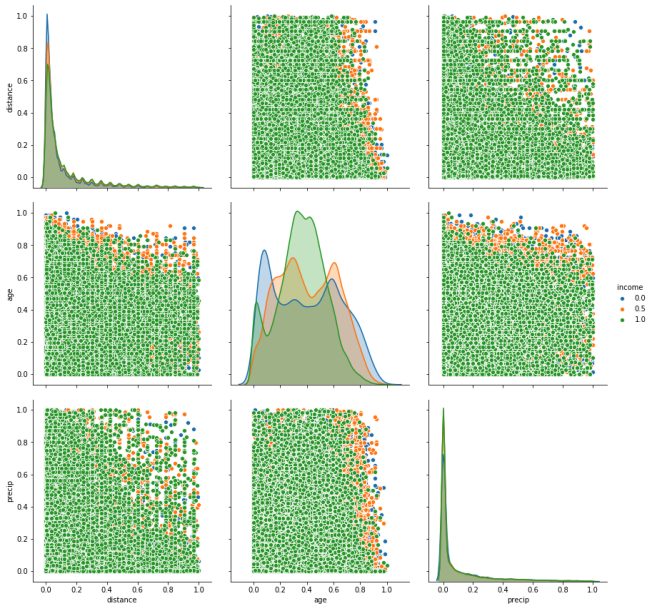


Figure 6: Distribución Características

4 CLUSTERING

Para la realización del *clustering* utilizando la técnica asignada *Fuzzy C-Means*, se ha hecho uso de la función *cmeans*, contenida en la librería *skfuzzy*. Con esta función se han utilizado una serie de parámetros necesarios para su ejecución: las características seleccionadas y el número de *clusters*, que se discutirán más adelante; el valor de m, que es un valor relacionado con la función de pertenencia (por defecto, 2); el error de distancia máximo; el número máximo de iteraciones y, por último, una semilla con el fin de reducir la aleatoriedad del algoritmo y obtener resultados deterministas en cada una de las diferentes ejecuciones.

Por otra parte, se ha decidido realizar la tarea de *clustering* seleccionando las características *distance*, *income* y *precip*, de manera que se puedan agrupar los viajes teniendo en cuenta la distancia de estos con respecto a los días lluviosos y según el nivel de ingresos de cada persona que lleva a cabo dicho viaje.

Antes de realizar el clustering, se han seleccionado las características indicadas anteriormente y se han agrupado según la característica *income*, con el fin de poder observar directamente cómo se distribuyen los datos (Figura 6) de manera rápida. De esta manera, observando los resultados, se puede ver que la mayoría de las personas con una edad media poseen una mayor cantidad de ingresos que, por ejemplo, las personas con menos y más edad, que se sitúan en el rango de ingresos medio-bajo. El resto de relaciones se encuentran representadas, mayoritariamente, de manera homogénea.

A la hora de realizar el *clustering*, se ha tenido en cuenta previamente el *Fuzzy Partition Coefficient* (FPC), que se trata de una métrica cuya finalidad es la de describir cómo de bien los datos se ajustan a un modelo determinado e indica el mejor número de *clusters* para los datos proporcionados. Se encuentra definido en un rango de entre 0 y 1, donde 1 representa un mejor valor. Se ha considerado la creación de entre 2 y 10 *clusters* para obtener los valores del FPC donde, a priori, se obtiene 2 como la mejor cantidad de *clusters* (Figura 7).

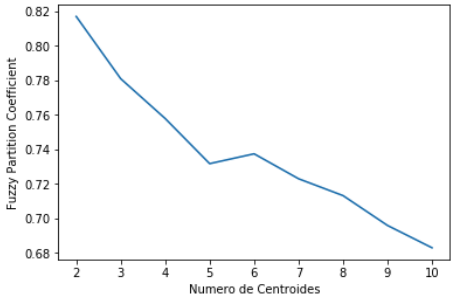


Figure 7: Gráfica FPC

A pesar de esto, si se observan los centroides de ambos *clusters*, se puede determinar que el algoritmo ha separado únicamente por la última de las características que, en este caso, es *income* (Figura 8). Por tanto, aunque esta es diferente en ambos *clusters*, tanto *precip* como *distance* toman valores muy similares y escoger 2 como el número de *clusters* no sería decisivo a la hora de realizar el agrupamiento.

```
[[0.10921362 0.08994534 0.97546108]
 [0.0880945  0.0870589  0.41742184]]
```

Figure 8: Centroides con 2 *clusters*

No obstante, la gráfica del FPC muestra un único pico cuando el número de *clusters* alcanza un número de 6, donde vuelve a bajar de nuevo para continuar la tendencia hasta el número de *clusters* considerados. En este caso, si observamos los centroides obtenidos, podrían mostrar una mayor diferencia entre algunos de ellos (Figura 9), lo que podría ser más beneficioso.

```
[[0.08488961 0.4602209  0.48999118]
 [0.48924842 0.04900858 0.96319436]
 [0.0568648  0.02621993 0.49983787]
 [0.08253558 0.50805717 0.97882619]
 [0.05489752 0.02735144 0.99777663]
 [0.05999601 0.04809536 0.00713131]]
```

Figure 9: Centroides con 6 *clusters*

Una vez que se han procesado y observado los resultados relativos a los *clusters*, se puede concluir con que algunos de ellos son similares, lo que podría confirmar el gráfico FPC obtenido anteriormente. Se han obtenido un total de 6 *clusters*, con un número diferente de trayectos (67.724, 14.671, 21.296, 64.744, 14.333 y 12.210). Con todos los datos disponibles como resultado, se han podido extraer algunas conclusiones:

- La mayoría de las personas con los mayores ingresos realizaron viajes de entre 0.1 km y 10 km. Esto se puede deducir de los *clusters* 3 y 5. Además, entre estas personas se pueden distinguir las que realizaron viajes con nula o escasa precipitación (0.61 mm, de media) y los que realizaron viajes con una mayor cantidad de precipitación (8.6 mm, de media). También, en comparación, los viajes cuando hubo una mayor precipitación fueron algo más largos. Por último, en este *cluster* se encuentran las personas con la mayor cantidad tanto de bicicletas como de coches, lo que podría indicar que estas personas pertenecen a una familia relativamente numerosa, en parte, debido al alto nivel de ingresos.
- La mayoría de las personas con ingresos menos elevados que el máximo (con una media de 2.8) fueron los que realizaron los viajes más largos. De media, estos viajes fueron de alrededor de 45 km de distancia, con un mínimo de 25 km. Esto se puede extraer del *cluster* 1.
- Por contra, las personas con los ingresos más bajos realizaron viajes con una distancia media de 7.27 km. Hasta el percentil 75 se encuentra una distancia de 7.2 km, lo que indica que este tipo de personas realizan los viajes más cortos, quedando representadas en el *cluster* 2. Estas personas también son las que cuentan con un número tanto de bicicletas como de coches menor que cualquier otro *cluster*, así como las que poseen el permiso de conducir, probablemente, por percibir una menor cantidad de ingresos.

5 CONTRASTE DE HIPÓTESIS

Por último, esta sección corresponde con el hito 2 del trabajo, donde se ha realizado un contraste de hipótesis con algunas de las características utilizadas en el *clustering*. Este contraste tiene como finalidad el hecho de demostrar cómo de representativas son las características utilizadas. Es decir, si los valores de estas han sido registrados al azar o, por ejemplo, han seguido una cierta distribución. Para ello, se ha realizado este contraste con las características *precip* y *distance*, con un valor de confianza (alpha) del 5%.

En ambos casos, la ejecución del contraste de hipótesis arroja como resultado un *p-value* de 0, por lo que se puede decir que se rechaza la hipótesis nula. De esta manera, por regla general, los datos utilizados para llevar a cabo la tarea de *clustering* no siguen una distribución concreta, siendo representativos y adecuados para ello.

6 CONCLUSIONES

Este trabajo ha servido para tratar y aplicar a un caso real algunos de los conceptos vistos y aprendidos en clase. En este caso, se ha lidiado con el método de *clustering Fuzzy C-Means*, junto con el tratamiento anterior necesario de todo el conjunto de datos, así como la importancia de otras técnicas como la estandarización y eliminación de *outliers*.

REFERENCIAS

[1] **Librería skfuzzy**. https://pythonhosted.org/scikit-fuzzy/api/skfuzzy.cluster.html#cmeans
[2] **Librería Pandas**. http://pandas.pydata.org/pandas-docs/stable/
[3] **SciPy - Kruskal**. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html
[4] Material de la asignatura de Desarrollo de Sistemas Inteligentes.