Trabajo de Clustering: Utilizando DBSCAN

Diseño de Sistemas Interactivos

Carlos Córdoba Ruiz
UCLM Ciudad Real
Carlos.Cordoba@alu.uclm.es

INTRODUCCIÓN

El objetivo de este trabajo es realizar un estudio exploratorio de los datos usando análisis clúster. Para ello se han de seleccionar las variables que sean relevantes, definir cuántos tipos de viajes distintos se pueden encontrar, analizar la importancia de cada grupo, analizar la correlación con las variables que los conforman, etc.

En específico, para este trabajo, se ha tenido que realizar todo el análisis clúster utilizando el algoritmo DBSCAN, que es un tipo de algoritmo basado en densidad.

El trabajo está conformado por dos Hitos y todo el código utilizado para realizar el trabajo podrá ser encontrado en este enlace de GitHub: https://github.com/CarlosCordoba96/Practica-Clustering.

1 HITO 1

El objetivo de este primer hito consiste en descubrir patrones de viajes en el conjunto de datos ofrecidos. Para ello se ha dividido en tres tareas.

1.1 Tarea 1

Los datos ofrecidos para el estudio se tratan de registros individuales de diarios de viajes de personas que participaron en una encuesta. Los datos que se recogieron de la encuesta fueron varios, y de distinto tipo.

Trip distance: Esta variable de los datos describe la distancia total del viaje en Km. Dentro de todo el conjunto de datos esta variable cuenta con el **valor mínimo** de 0.1 KM, un **valor máximo** de 400 KM, una **Media** de 12.218 y una **Desviación estándar** 23.546 KM.

Weekend: Esta variable significa si un viaje fue realizado en fin de semana. Puede tener el valor **yes** o **no**. También cabe mencionar que el 82.066% de los datos tienen el valor **no**, lo que quiere decir que el 82.066% de los datos no fueron tomados en fin de semana.

Mode: Describe el modo de transporte en el que se realize el viaje. Esta variable puede tener cuatro valores distntitos, **walk** (20.935% de los datos), **bike** (24.473%), **pt** (transporte público, 2.316%) y **car** (52.276%).

Individual age: La edad de los participantes, medida en años, el **mínimo** que se ha registrado en los datos son de 18 años y el **valor máximo** son 98 años, la **media** de edad es de 47.661 y la **desviación estandar** 15.935.

Education: La educación del participante, puede tener tres valores distintos: **low** (27.370%), **middle** (38.293%) y **high** (34.337%).

Ethnicity: Etnia del participante, puede tener tres valores, **native** (87.404%), **western** (7.07%) y **other** (4.889%). La gran mayoría de los participantes de la encuesta son personas nativas.

License: Indica si la persona a la que se le ha realizado la encuesta dispone de licencia de conducer o no. Puede tener el valor **No (**10.243%) y **Si**. Cómo se puede observer la gran mayoría de personas a las que se le realize la encuesta disponen de carnet de conducer.

Male: Si el participante de la encuesta es hombre. Puede tener el valor **No** (54.498%). Se puede decir que el 54.498% de los participantes son mujeres, el resto son hombres.

Bicycles: Número de bicicletas por casa. El **mínimo** es 0 y el **máximo** es 10, la **media** de todos los registros es de 3.357, y la **desviación estándar** 1.937.

Cars: Número de coches por casa. El **mínimo** es 0 y el **máximo** 10, la **media** de coches es de 1.383 de todos los participantes, y la **desviación estándar** es 0.822.

Income: Ingreso neto anual de cada casa expresado en miles de euros. Puede tener tres valores distintos, el primero **<20** (11.832%), **>=20-40** (42.123%) y **>=40** (46.044%).

Density: Densidad de direcciones, expresado en 1000 direcciones por km cuadrado. El **valor mínimo** es de 0.002 y el **valor máximo** es de 11.443, el **valor de la media** es de 1.569 y la **desviación estándar** 1.593.

Diversity: Índice de diversidad de Shannon de clases de uso de la tierra. El **valor mínimo** es 0 y el **valor máximo** 2.828, la **media** de todos los registros es 1.775 y la **desviación estándar** 0.493.

Green: Proporción de espacio verde por codigo postal en tanto porciento. El valor **medio** es 0 y el valor **máximo** 97.813, la **media** es 54.939 y **la desviación estándar** 22.172.

Precip: Cantidad de precipitación en mm. El **valor mínimo** es 0 y el valor **máximo** 132.3, la **media** 2.185 y **la desviación estándar** 4.675.

Temp: Temperatura máxima en ºC. La temperatura **minima** es de -9 y la **máxima** 35.9, la **media** de todos los registros es 13.317 y la **desviación estándar** 22.172.

Wind: Media de la velodidad del aire en m/s. El valor **mínimo** es de 0.400 y el **máximo** 16.3, la **media** 4.098 y la **desviación estándar** 1.915.

1.2 Tarea 2

La segunda tarea que se realizó fue realizar un análisis clúster de los datos para determinar los patrones de viajes. Antes de aplicar dicho algoritmo se debían tratar los datos, ya que había variables categóricas, no numéricas, que para ser tratadas en los algoritmos de Clustering se debían transformar a numéricas, son las variables: "male, ethnicity, mode_main, education, income, license, weekend". Es por ello que para las variables que definían una categoría o valor único se decidió cambiarles el valor por números enteros, por ejemplo, la variable **"mode_main"** tiene cuatro posibles valores diferentes, "walk, car, bike, pt", entonces se le asignaron a cada uno un valor entero desde 0 hasta 3. Lo mismo con la variable **"male"**, los valores podían ser "yes o no" por lo tanto se le asigno "1" cuando fuera "yes" y "0" en caso contrario.

De esta misma forma se trataron a las variables: **"weekend", "license", "ethnicity"**

Pero había dos variables que se referían a intervalos, **"income"** y **"education"**, en estas variables se decidió un incremento de 0.5 entre cada uno de los valores para poder tratar los datos. Por lo tanto en el ejemplo de la variable **"income"**, cuando tuviera el valor "less20" se le asignaría el valor 0, cuando fuera "20to40" el valor 0.5 y cuando fuera "more40" el valor 1. Y en el caso de **"Education"** cuando fuera "lower" sería 0, en el caso de "middle" 0.5 y en el caso de "higher" 1.

Una vez se trataron las variables el siguiente paso era tratar el problema de las magnitudes de las variables, ya que por ejemplo la magnitud de la edad que tiene un rango totalmente distinto a la variable de la distancia, para ello se decidió utilizar un preproceso de los datos utilizando el escalador Min/Max.

El problema para realizar la parametrización de los datos que al realizar el pairwise con 230,608 registros la memoria del ordenador fallaba todo el rato, es por eso que se decidió extraer 20.000 cómo una muestra para realizar dicho estudio.

Para proceder al estudio de los datos se decidió eliminar la columna de la variable "diversity" y "ethnicity" para proceder a estudiar todos los clusters de la muestra.

Para hacernos una idea de los datos que tenemos se ha ejecutado el PCA y graficados nuestros datos para ver las agrupaciones de los datos para intentar extraer información antes de aplicar el algoritmo de Clustering, a continuación, se muestra el resultado:

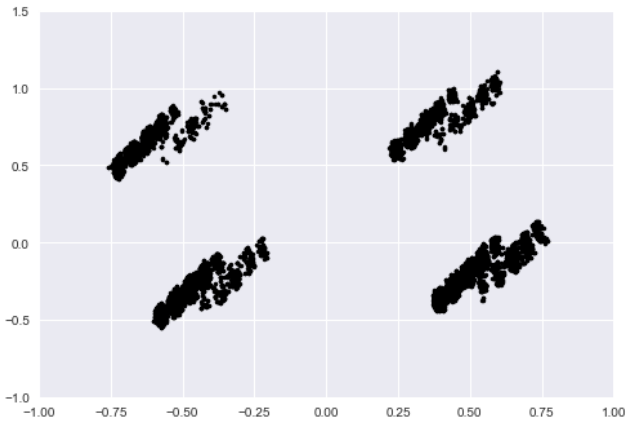

Figura 1- PCA

A primera vista se pueden observar cuatro regiones densas de datos, lo que puede corresponder con cuatro clusters diferentes.

A la hora de ejecutar el método DBSCAN se piden dos parámetros: el primero **MinPts:** el cual se trata del número mínimo de puntos requeridos para que una región se considere densa, al tratarse de una muestra de 20.000 registros se ha considerado que una región pueda considerase densa sea con 500 registros. Una vez se tiene establecido el MinPts hay otra variable que se debe especificar, **Eps**: el número de puntos en un radio específico, para elegir este parámetro se ha realizado una gráfica en la que se ordena de menor a mayor la distancia de todos los puntos a su k-ésimo vecino, y el valor en el que se produce un cambio de la curva es un buen valor para Eps, por ello se ha realizado dicha gráfica y estudiado el coeficiente de Silhouette para ver que valor de Eps puede ser mejor.

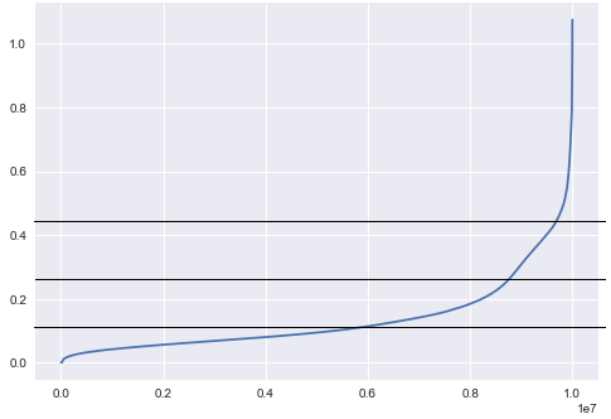

Figura 2- K-vecinos

Se puede ver que en la figura que el cambio que se produce no es muy brusco, por eso se ven dos intervalos, dónde se ve un pequeño crecimiento, y se pueden mostrar los valores que van a ser estudiados, principalmente el superior a este, aproximadamente desde 0.3 a 0.42. El valor seleccionado ha sido 0.4, ya que ha conseguido un coeficiente de 0.262. A continuación se muestra una figura de los datos de la muestra organizado en función de los

clusters encontrados utilizando el algoritmo DBSCAN.

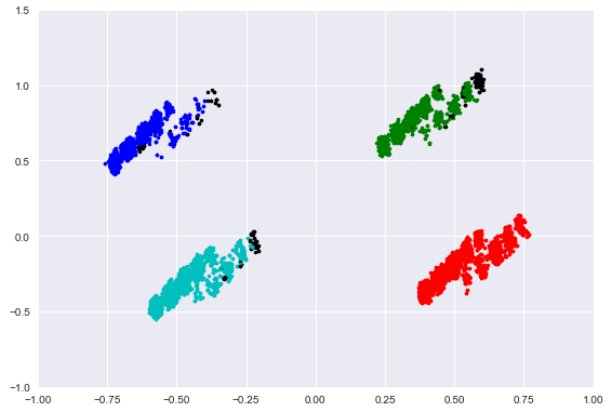

Figura 3- Clusters

Se pueden observar cinco clusters, los cuales cuatro de ellos representados con los colores y otro que se puede corresponder con algunos outliers de cada uno de los grupos representados con color negros.

Hasta este paso se puede observar todo lo realizado en el script "Practica.py" situado en el repositorio Github mencionado al comienzo del documento.

1.3 Tarea 3

En esta última tarea del primer Hito se ha realizado un análisis descriptivo de cada uno de los clusters definidos en la tarea dos, a partir del algoritmo DBSCAN.

Se han obtenido cinco clusters, vamos a analizar cada uno de ellos:

Cluster 0: Corresponde con la nube de datos color azul oscuro. Con un número de 2637 registros, de ellos, 1594 corresponen a viajes realizados en coche, 604 a viajes realizados andando, 398, en bici y 41 en transporte público, una peculiaridad de este cluster es que todos los datos pertenecientes a este cluster son **únicamente hombre**s y también los viajes fueron realizados en **el fin de semana**. La media de la **distancia** es de 16.1322 KM con una desviación estándar de 28.9482. La media de la **densidad** tiene el valor de 1.53, la media de la **edad** es de 48.29 con una desviación estándar de 16.53. La media del **número de coches** es 1.45 con una desviación estándar de 0.8. La media de la variable **green** es 54.48 con una desviación estándar de 21.37. La **temperatura media** era de 11.77 con una desviación estándar de 8.92. La media de las **precipitaciones** es 1.97 con std de 4.25 y la media de la **velocidad del viento** es de 3.68 con std de 1.46. De todos los registros únicamente 115 no dispone de licencia de conducir.

Cluster 1: Corresponde con la nube de datos de color verde. Con un número de 2698 registros. Al igual que pasaba en el Cluster número 1, en este cluster los 2698 registros que tiene, **todos ellos son mujeres** y además también todos los viajes se realizaron **en fin de semana**. La media de la **distancia** de este grupo es de 14.66 con una std de 27.26. Una **densidad** media de 1.59 y una std de 1.66. La **edad** media de la gente de este cluster es de 46.63 y un std de 15.30. La media del número de **coches** es de 1.39 y un std de 0.75. La variable **green** tiene una media de 53.04 y una std de 20.62. La **Temperatura** media es de 11.69 con una std de 8.49. La media de las **precipitaciones** es de 2.40con una std de 4.98. La media de la **velocidad** del aire es 3.77 con una std de 1.5. De todos los viajes 1657 se realizaron en coche, 661 andando, 340 en bicicleta y 40 en transporte público. De todos estos solamente 320 personas no tienen la licencia de conducir.

Cluster 2: Corresponde con la nube de datos de color rojo. Con un número de 7982 registros. Este grupo está formado únicamente por **mujeres** y además los viajes realizados **no han sido realizados en fin de semana**. De estos viajes, 1612 han sido realizados **andando**, 3826, en **coche**, 2279 en **bicicleta** y 265 en **transporte público**. 2353 tienen una **educación baja**, 2881 una **educación media** y 2748 una **educación alta**. Se puede observar que en la variable **income** únicamente 949 disponen de less20, 3273 20to40 y 3760 more40. De todos los registros 1076 no tienen **licencia** de conducir y 6906 si lo tienen. La media de las **distancias** son 7.58 con una std 15.06. La media de la variable **densidad** es 1.56 con una std de 1.51. La media de **edad** es 46.09 con una std de 14.65. La media del número de **coches** por casa es 1.35 con una std de 0.77. La media de las **bicicletas** por casa es 3.48 con una std de 1.96. La media de la variable **green** dentro del grupo es 54.39 y una std de 22.54. La media de las **precipitaciones** son 2.12 con una std de 4.48. La media de la **velocidad del viento** es 3.89 con una std de 1.47. La media de la temperatura 10.46 con una std de 8.78.

Cluster 3: Corresponde con la nube de datos de color cyan. Con un número de 6412 registros. En este grupo se da la peculiaridad de que todos los registros pertenecen a **hombres** y que los viajes **no fueron realizados en el fin de semana**. Dentro de ese grupo, 946 de los viajes realizados fueron **andando**, 3881en **coche**, 1358 en **bicicleta** y 227 en **transporte público**. Los **ingresos** de este grupo, 543 registros son **menos que 20**, 2783 son **entre 20 y 40** y 3086 son **mas de 40**. Solamente 332 de los registros, no tienen **licencia** de conducir. La **distancia** media de este grupo es 15.52 con un std de 26.54 con hasta un máximo de 280. La **densidad** media es 1.51 con un std 1.57. La **edad** media está en 48.21 con un std de 16.13. La media del **número de coches por casa** es de 1.46 con un std de 0.82. El número medio de **bicicletas** por casa es 3.40 con un std de 1.95. La medida media de la variable **green** es 56.46 con un std de 22.25. La **temperatura** media es de 10.29 con un std de 8.67. La media de las **precipitaciones** es de 2.26 con un std de 4.82. La media de la **velocidad del viento** es de 3.86 con un std de 1.46.

Cluster -1: Corresponde con los datos de color negro, considerados outliers del resto de clusters. Con un numero de 271 registros. Aproximadamente contiene la mitad de sus datos **hombres** y mujeres, 137 de mujeres y 134 hombres. Solamente 70 viajes del grupo fueron realizados fuera del **fin de semana**, los 201 restantes en fin de semana. Según la variable **income**, hay 265 registros de personas con el valor **less20** y 58 **20to40**. La **distancia**

media son 12.834 con un std de 27.88. La **densidad** media es 3.11 con un std de 2.6. La **edad** media del grupo es 50.74 con un std de 21.81. El número medio de **coches** es 0.15 con un std de 0.39. La media del número de **bicicletas** es 1.47 con un std de 1.10. La media de la variable **green** es 37.22 con un std de 22.44. La media de la **temperatura** es 11.82 y un std de 8.89. La media de las **precipitaciones** es 2.60 con un std de 5.57. La media de la **velocidad del viento** es 3.91 con un std de 1.68.

Cómo primeras conclusiones se pueden sacar que principalmente de los clusters 0 a 3, las principales variables son "male" y "weekend" ya que tienen una gran separación de los grupos en función de su valor. Por otro lado, el cluster -1 es un conjunto de datos que no corresponden a ninguno de los otros clusters y se caracterizan principalmente que la mayoría de los registros fueran en fin de semana, el número medio de coches por casa es bastante bajo y los grupos son de registros de persona que los ingresos son menos de 20 y otros poco entre 20 y 40.

Para realizar esta tarea se creó el script AnalyseData.py disponible en el repositorio de GitHub. Además, en el repositorio de GitHub se han añadido unas gráficas con las medias de cada variable de cada cluster para visualizarlas.

2 HITO 2

En este Segundo hito se tratará de realizer una descripción semántica de los grupos.

Para ello se ha procedido a utilizar el test de Kruskal-Wallis para testar que variables son importantes para diferenciar un grupo de otro. Para este hito se ha creado el script Hito2.py disponible en el repositorio GitHub.

La primera variable que se ha analizado ha sido la variable **"male"** la cual se ha rechazado la hipótesis nula, por lo que se sabe que esta variable toma importancia para diferenciar los grupos, lo mismo ha pasado con la variable **"weekend"**, por lo que se puede asumir que las conclusions que sacamos sobre estas variables en el hito2 son correctas.

Respecto la variable **"density"** se han hecho una serie de análisis y entre el cluster 0 y el cluster -1 se rechaza la hipótesis, por lo que sería significante esta variable, pero con el estudio de los clusters 0, 1 y 2 se falla a rechazar la hipótesis, por lo tanto no es una variable que se diferencie en esos grupos, pero al analizar los cluster 0,1,2 y 3 si se rechaza la hipótesis, por lo que podría ser importante esa variable para discerner si pertenece en ese cluster o no.

Respecto la variable **"precip"**, entre los clusters 0 y 1 y 0 y -1 se falla al rechazar la hipótesis, por lo tanto tienen unas distribuciones parecidas, pero con los clusters 2 y 3 se rechaza la hipótesis por lo tanto en los clusters 2 y 3 tienen distribuciones distintas respecto a 0 y -1, y además también se rechaza la hipotesis entre esos grupos, por lo tanto es una variable importante para diferenciar grupos.

Respecto la variable **"green"**, se ha fallado a rechazar la hipótesis entre los grupos 0 y 2, lo que quiere decir que la variable es parecida en estos grupos, el análisis con todos los clusters se ha rechazado, por lo tanto es otra variable interesante para tomar en cuenta.