Trabajo de Clustering: Utilizando DBSCAN

Diseño de Sistemas Interactivos

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

ABSTRACT

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 qu

1.1 Tarea 1

LOS DATOS

The **DisplayFormulaUnnum** style is applied only in case of an unnumbered equation. An unnumbered display equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.



Figure 1: Figure Caption and Image above the caption [In draft mode, Image will not appear on the screen]

**Theorem/Proof/Lemma.** Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement.

....Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract.

1.2 Tarea 2

In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

1.2 Tarea 2

La segunda tarea que se realizó fue realizar un análisis cluster 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”**

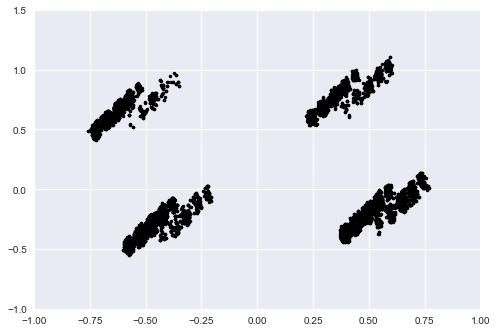
Pero había dos variables que se referían a intervalos, **“income”** y **“education”**, en estas variables se decidió un incremento de 0.5 entre cada uno de los valores para poder tratar los datos. Por lo tanto en el ejemplo de la variable **“income”**, cuando tuviera el valor “less20” se le asignaría 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.

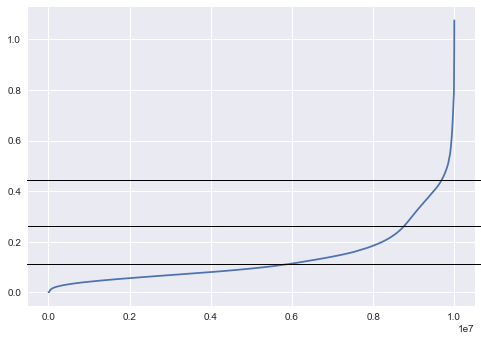
Para proceder al estudio de los datos se decidió eliminar la columna de la variable “diversity” y “ethnicity” para proceder a estudiar todos los clusters de la muestra.

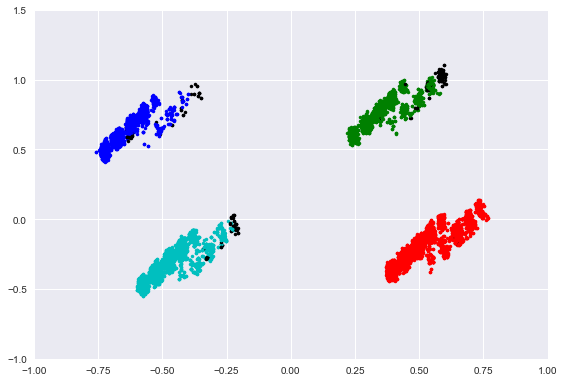
Para hacernos una idea de los datos que tenemos se ha ejecutado el PCA y graficados nuestros datos para ver las agrupaciones de los datos para intentar extraer información antes de aplicar el algoritmo de Clustering, a continuación, se muestra el resultado:



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.



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.

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 .
* **Cluster 1:** Corresponde con la nube de datos de color verde.
* **Cluster 2:** Corresponde con la nube de datos de color rojo.
* **Cluster 3:** Corresponde con la nube de datos de color cyan.
* **Cluster -**1: Corresponde con los datos de color negro, considerados outliers del resto de clusters.

are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

*1.1.1 Heading Level 3.* Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

*1.1.1.1 Heading Level 4.*Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10.

[2] F.N.M Surname, Article Title, https://www.acm.org/publications/proceedings-template.

[3] F.N.M Surname and F.N.M Surname, 2018 Article Title, The title of book two (2nd. ed.). Publisher Name, City, State, Country.

[4] Ian Editor (Ed.). 2018. *The title of book two* (2nd. ed.). University of XXX Press, City, Chapter 100. DOI: <http://dx.doi.org/10.1000/0-000-00000-0>.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00