IMPORTANCIA DE LA MÉTRICA DE DISTANCIA EN LOS ALGORITMOS DE CLUSTERING

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

Emanuel Ciardullo, estudiante Lic. En Estadística

INTRODUCCIÓN

En estadística se conoce como análisis clúster a los métodos que comparten el objetivo de identificar grupos homogéneos dentro de un conjunto de datos. Para ello, los individuos pertenecientes a un mismo grupo deben ser lo mas similares posible mientras que a su vez los grupos deben ser disimiles entre si. Es por esto, que para realizar el agrupamiento es indispensable definir alguna medida de similaridad entre los individuos.

Las métricas de distancia cuantifican la similaridad entre individuos de a pares, a partir de características cuantitativas y/o cualitativas. A lo largo de los años se han derivado varias métricas de distancia y si bien no existe una métrica de distancia óptima para todos los casos, se puede encontrar una métrica que funcione mejor que el resto para un conjunto particular de datos.

En este trabajo se evalúan 4 medidas diferentes en un estudio de morfología de especies de flores.

OBJETIVO

Evaluar las diferencias en los agrupamientos obtenidos a partir de un algoritmo de clasificación (*k-medoid*) aplicados sobre un conjunto de datos de un problema real, cuando varia la forma de definir las distancias entre observaciones, considerando únicamente métricas para variables cuantitativas.

BASE DE DATOS

Se utiliza la base de datos Iris introducida por Fisher en 1936. Es un conjunto de datos multivariados que contiene 50 muestras de cada una de tres especies de Iris: Iris setosa (C_1), Iris virginica (C_2) e Iris versicolor (C_3). Se midieron cuatro atributos de cada muestra: el largo y ancho del sépalo y pétalo, en centímetros. Se agrega para cada unidad un indicador de la población de pertenencia. Los datos fueron estandarizados antes de la aplicación de *K-medoid*.

ALGORITMO DE CLUSTERING

Se utiliza el algoritmo *K-medoid.*

1) Se eligen K puntos iniciales aleatoriamente $(m_1, m_2, \dots, m_k)$.

2) Se asignan todos los puntos de la base de datos al *medoid* más cercano.

3) Para cada uno de los K clústeres se comprueba si no existe otra observación que tomada como *medoid* consiga reducir la distancia promedio del clúster.

$$distancia\ promedio\ \ \overline{d_k} = \frac{\sum_{i=1}^{n_k} d(x_i,\ m_k)}{n_k}$$

Según la función de distancia que se haya escogido.

4) Si esto ocurre se selecciona a la observación que consigue una mayor reducción de la distancia promedio como nuevo *medoid*.

5) Si al menos un *medoid* se modificó en el paso anterior volver al paso 2, sino se termina el algoritmo.

VALIDACIÓN

Sean k clases $(C_1, C_2, \dots, C_k)$ en el conjunto D de N observaciones. El algoritmo de agrupamiento produce k clústeres $D_1, D_2, \dots, D_k$ de $n_1, n_2, \dots, n_k$ datos respectivamente.
Las medidas utilizadas para evaluar la consistencia de los resultados de *K-medoid* al variar las métricas de distancia son entropía y pureza.

ENTROPÍA: Es una medida del desorden en los clústeres. Para el clúster i con $i = \overline{1,k}$, se define:

$E_i = -\sum_{j=1}^{k} \Pr_i(C_j) * \log_2\left(\Pr_i(C_j)\right)$,

donde $\Pr_i(C_j)$ es la proporción de puntos de la clase C_j ubicados en el clúster D_i.

La entropía total, es: $E = \sum_{j=1}^{k} E_i * {}^{n_i}/_{N}$

A menor entropía mejor es la calidad de la clasificación.

PUREZA: Cuantifica si los clústeres contienen solo una calase entre sus datos, o mas. Para el clúster i con $i = \overline{1,k}$, se define: $U_i = \max_j\left(\Pr_i(C_j)\right)$

La pureza total, es: $U = \sum_{i=1}^{k} U_i * {}^{n_i}/_{N}$

A mayor pureza mejor es la calidad de la clasificación.

RESULTADOS

La distancia entre dos puntos $x = (x_1, x_2, \dots, x_n)$ e $y = (y_1, y_2, \dots, y_n)$

CANBERRA

$$d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Valores con numerador o den. 0 se eliminan de la suma.

K-medoid	Grupos reales C_1	C_2	C_3
$\widehat{C_1}$	50	4	0
$\widehat{C_2}$	0	25	1
$\widehat{C_3}$	0	21	49

Pureza: 0,83
Entropía: 0,49

EUCLIDEA

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

K-medoid	Grupos reales C_1	C_2	C_3
$\widehat{C_1}$	50	0	0
$\widehat{C_2}$	0	41	14
$\widehat{C_3}$	0	9	36

Pureza: 0,85
Entropía: 0,51

PEARSON

$$d(x,y) = 1 - \frac{\sum_{i=1}^{n} x_i\, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

K-medoid	Grupos reales C_1	C_2	C_3
$\widehat{C_1}$	49	2	0
$\widehat{C_2}$	0	16	13
$\widehat{C_3}$	1	32	37

Pureza: 0,68
Entropía: 0,77

MAXIMA

$$d(x,y) = \max_i(x_i - y_i)$$

K-medoid	Grupos reales C_1	C_2	C_3
$\widehat{C_1}$	49	0	0
$\widehat{C_2}$	1	42	19
$\widehat{C_3}$	0	19	31

Pureza: 0,81
Entropía: 0,58

CONCLUSIONES

Los valores obtenidos en los índices utilizados para la validación de los clústeres denotan la variabilidad que se puede obtener en los resultados de un mismo algoritmo de clustering a partir de únicamente variar la métrica de distancia. En este caso ,que se consideraron cuatro métricas distintas, se observó una variación máxima de 0,17 en la pureza de los clústeres y de 0,28 en la entropía. Esta variación en los resultados también es observable en las tablas cruzadas entre las etiquetas a priori que se tenían para los datos y las obtenidas por *K-medoid*. Por lo tanto, en cada caso particular la elección de la métrica es un tema de importancia. La recomendación para los analistas es que sean evaluadas distintas posibilidades dentro de un conjunto de métricas adecuadas a la naturaleza y estructura de los datos para decidir a posteriori cual de ellas conduce a un agrupamiento de mejor calidad.