Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO III: Minería de datos

PARTE 02: Clustering (K-Medias y algoritmos jerárquicos)

Introducción:

En este trabajo se abordarán los algoritmos de agrupamiento o *clustering* y las medidas de distancia asociadas a efectos de determinar la similitud de los datos.

En primer lugar, se trabajará con algunas de las medidas de distancia clásicas para variables numéricas como la euclídea, Manhattan y la distancia de Minkowski y otras medidas relacionadas a calcular distancias en variables binarias y categóricas.

Luego, se utilizarán los software **R** y **Weka** con el objetivo de resolver problemas de la disciplina, los cuales son una combinación ejercicios clásicos de minería de datos complementados con ejercicios propuestos por el equipo docente.

Consignas:

1. **Medidas de distancia.** Calcule la distancia entre los siguientes puntos y el centroide (2, 4) utilizando las medidas euclídea, Manhattan y Minkowski (con p = 3):

X	Y
4	8
9	17
3	7

   ¿Encuentra diferencias relativas entre las diferentes métricas utilizadas y el resultado obtenido? Explique el comportamiento de cada una utilizando gráficas.

2. A continuación, calcule la distancia entre las diferentes variables de tipo categóricas con respecto a la instancia {1, lluvioso, templado, alta, fuerte}:

#	PRONÓSTICO	TEMPERATURA	HUMEDAD	VIENTO
2	lluvioso	Frío	normal	fuerte
3	nublado	Frío	normal	fuerte
4	soleado	templado	alta	leve
5	soleado	Frío	normal	leve
6	lluvioso	templado	normal	leve
7	soleado	templado	normal	fuerte
8	nublado	templado	alta	fuerte
9	nublado	Calor	normal	leve
10	lluvioso	templado	alta	fuerte

¿Cuáles son las instancias más cercanas a la instancia #1?

3. Ahora, y a partir de los datos de la siguiente tabla, agrupe los datos de acuerdo al algoritmo k-medias utilizando la medida *euclídea* y con los puntos A1, A2 y A7 como centroides iniciales.

PUNTO	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	2	7
A5	7	5
A6	6	4
A7	1	2
A8	4	9

4. Implemente en un lenguaje de programación el algoritmo **k-medias** de acuerdo a alguna de las métricas de distancia vistas antes.

5. **K-means.** Se provee un dataset[1] sobre las características internas del núcleo de tres clases de trigo diferentes. Cargue el dataset en una de las herramientas de minería de datos provistas y resuelva:
    a. Utilice el algoritmo k-medias variando la cantidad de centroides a efectos de agrupar los datos de la manera más eficiente.
    b. ¿Cuál es la cantidad de grupos que permite un mejor agrupamiento de los datos? ¿Mediante cual métrica puede verificar esto?
    c. ¿Cuáles son las características más distintivas de cada uno de los cluters resultantes?
6. Ahora, trabaje sobre el dataset *ds-abandono.xls*:
    a. Analice y describa las características más salientes de cada uno de los grupos encontrados por el algoritmo. ¿La información provista responde a criterios lógicos? ¿En qué casos?
    b. Encuentre la cantidad de grupos que logran el mejor agrupamiento para los datos. Justifique la elección a partir de métricas y gráficas de los conglomerados resultantes.
    c. Ahora aplique algún algoritmo jerárquico a efectos de agrupar los datos. ¿Cuál nivel se corresponde con el agrupamiento realizado por k-medias en el **punto 6) a)**?
    d. ¿El agrupamiento jerárquico permite encontrar una mejor forma de agrupar los datos? Si fuera así, ¿Cuál es ese agrupamiento?
7. **Algoritmos jerárquicos[2] (hclust y diana).** Incorpore en R nuevamente el dataset del punto 5 y realice las siguientes actividades:
    a. **Métodos aglomerativos.** Realice el agrupamiento de los datos a través del método aglomerativo. Tenga en cuenta la función *hclust().*
    b. Grafique el resultado y escoja cual es el nivel que mejor agrupa los datos.

[1] Disponible en https://archive.ics.uci.edu/ml/datasets/seeds
[2] Para trabajar con R utilizando las funciones relacionadas a estos algoritmos previamente deberá instalar la librería **cluster.**

c. **Métodos divisivos.** Realice el agrupamiento de los datos a través del método divisivo. Tenga en cuenta la función *diana().*
d. Grafique el resultado y escoja cual es el nivel que mejor agrupa los datos.
e. **Comparación.** Explique someramente la diferencia entre ambos métodos y compare los resultados encontrados en términos de:
    I. Los algoritmos,
    II. Las características de los clusters obtenidos.

Referencias sugeridas:

Data Mining: Practical Machine Learning Tools and Techniques

http://www.cs.waikato.ac.nz/ml/weka/book.html

Cluster Analysis (R)

http://www.statmethods.net/advstats/cluster.html

Data Mining: Concepts and Techniques. Jiawei Han & Micheline Kamber. Morgan Kaufmann. Second Edition. 2006. Chapter 7.