MINERIA DE DATOS UTILIZANDO SISTEMAS INTELIGENTES

PRACTICA 5 – AGRUPAMIENTO (CLUSTERING)

Material de Lectura: Capítulo 16 del Libro Introducción a la Minería de Datos de Hernández Orallo

Ejercicio 1

*Nota: para realizar esta práctica deberá instalar el operador Silhouette copiando el archivo **CPPlugin-0.3.jar** al directorio **lib/plugins** dentro del directorio de instalación de Rapidminer. Si tiene Rapidminer abierto, deberá reiniciarlo para que el operador pueda ser utilizado.*

El archivo **2d_simple.csv** contiene 5000 filas, cada una con dos atributos numéricos, **x,y,** junto con el cluster original. Estos datos son sintéticos (generados artificialmente).

1. Cargar el conjunto de datos en Rapidminer. Recordar que el atributo *cluster* debe ser configurado como *label* para que no se considere al hacer el clustering.

2. Determinar visualmente la cantidad de clusters utilizando algún gráfico de Rapidminer.

3. Utilizando los operadores **Clustering (K-Means)** y **Performance: (Average) Silhouette**, realice un clustering de los datos y evalúe el resultado con el índice promedio Silhouette. El clustering debe realizarse con **k=2** y distancia **Euclídea**.

4. Repita el punto 3 variando el valor de **k** (al menos con k=2,3,..,9)**.** Haga una tabla con el valor del índice Silhouette para cada valor de **k**. Con esa tabla, determine el valor de **k** óptimo ¿coincide con el valor que determinó en el punto 2? Recuerde que en el índice Silhouette es mejor si el valor es más alto.

5. Repita el punto 4, pero ahora con el puntuando el resultado del clustering con Davies-Bouldin en lugar de Silhouette. Para ello utilizar el operador **Performance (Cluster Distance Performance)**, con *main criterion="Davies Bouldin"*, y tildar "main criterion only", "normalize" y "maximize". Recuerde que el criterio Davies Bouldin es mejor cuando el valor es más chico.

6. Repita los puntos 1-5 para el conjunto de datos **2d_complex.csv.** Explicar por qué en este conjunto de datos el valor óptimo de k calculado con los índices **Silhouette y Davies Bouldin** no coincide con el óptimo determinado visualmente.

Ejercicio 2

El archivo **Vidrios2.csv** contiene 214 muestras que corresponden a propiedades físicas de diferentes tipos de vidrios. Cada una está compuesta por los siguientes atributos:

- RI: Indice de refracción
- Na: Porcentaje de sodiopresente
- Mg: Porcentaje de magnesio presente
- Al: Porcentaje de aluminio presente
- Si: Porcentaje de silicio presente
- K: Porcentaje de potasio presente
- Ca: Porcentaje de calciopresente
- Ba: Porcentaje de bario presente
- Fe: Porcentaje de hierro presente
- Type: Tipo de vidrio

Utilice RapidMiner para agrupar los datos utilizando el algoritmo K-medias. Pruebe diferentes métricas (Euclídea, Manhattan y Chebychev) y distintos valores de K.

A partir del mejor agrupamiento conseguido, evaluado por Davies-Bouldin o Silhouette:

a) Observe el valor del atributo “type” de los vidrios que componen cada uno de los clusters. ¿Qué puede decir al respecto? Recuerde que debió setear como “label” el atributo “type” para no incluirlo en el agrupamiento.

b) Analizando los valores de los centroides para cada uno de los atributos,¿cómo podría caracterizar los grupos conseguidos? Tenga en cuenta características en común o diferencias entre ellos.

Ejercicio 3

El archivo **Zoo.xls** contiene las características de 101 animales entre los que hay anfibios, aves, mamíferos, peces, insectos, invertebrados y reptiles. Marque como “id” el atributo “animal” y como “label” el atributo “clase”. Utilice RapidMiner para agrupar los datos utilizando el algoritmo K-medias.

Pruebe diferentes métricas (Euclídea, Manhattan, Chebychev, etc.) y distintos valores de K. Evalúe los resultados obtenidos en cada caso con los índices Davies-Bouldin y Silhouette.

Observe cómo fueron repartidos los animales en cada uno de los clusters prestando atención a los atributos “animal” y “clase” no tenidos en cuenta al momento de agrupar.