Ing. de Sistemas e Informática
Estructuras de Datos y Algoritmos
Jorge Mario Londoño P.

Taller 3

Estructura Unión-Búsqueda

Clusterización por el vecino más cercano

Los métodos de clusterización en análitica de datos buscan agrupar conjuntos de puntos "similares". Para efectos del ejercicio, se toma un conjunto de puntos 2D que el programa lee de un archivo plano separado por comas, donde cada linea representa las coordenadas de un punto.

El agrupamiento por vecinos cercanos se implementa de la siguiente forma:

1. Cada punto del dataset es inicialmente un cluster con un elemento.

2. Se busca la pareja de puntos más cercanos que no se encuentren "conectados" y se "unen".

3. Se repite el paso 2 hasta no encontrar pares de puntos con una distancia menor a DMAX (parámetro del algoritmo. Se deja como una constante estática en el programa).

El programa debe arrojar las siguientes salidas:

Un arreglo Nx1 indicando el cluster al que pertenece cada punto.

El número de clusters encontrados.

Para cluster indicar cuantos puntos contiene.

Ejercicios a desarrollar

1. Implementar el algoritmo de clusterización por vecino más cercano como una función de biblioteca. Utilizar ejemplos de archivos de puntos que se indican al final.
2. El método main de la clase tomar el nombre del archivo a leer de `args[0]`.
3. Graficar los puntos del cluster asignando distintos colores a los puntos de cada uno de los clusters.
4. Estimar análiticamente el desempeño del algoritmo en función del número de puntos N.
5. Evaluar empíricamente el desempeño del algoritmo. Incluir tabla de mediciones en función de N, gráfica y curva de mejor ajuste. Comparar el resultado con el valor obtenido analíticamente.

Entregables

Remitir el código fuente de la solución implementada y la hoja de cálculo con los resultados de las pruebas experimentales (se aceptan Word, Excel, LibreOffice, LibreOffice, PDF). Nombrar el archivo comprimido `Taller3-<Nombre1>-<Nombre2>…(.zip .rar .7z o .tgz)`. Para estandarizar la forma de invocar el programa, ubicar el método main y las funciones de biblioteca solicitadas en la clase `Taller3`.

En caso de utilizar estructuras de las bibliotecas del texto (`algs4.jar`) **no** anexar la biblioteca.

Grupos máximo de 2 personas.

Puntos de prueba

Archivos con puntos de prueba para el ejercicio:

- datapoints-k=2-n=200.csv (caso de prueba con 2 clusters bien separados)
- datapoints-100.csv
- datapoints-120.csv
- datapoints-150.csv
- datapoints-1000.csv
- datapoints-2500.csv
- datapoints-5000.csv

Estos archivos contiene puntos en el rango (-2,-2) a (2,2). Usar distancias máximas del orden de DMAX en el rango 0.1 - 0.3 aproximadamente.