Práctica 3 – Computación Distribuida y Escalable con Hadoop

Cloud Computing: Servicios y Aplicaciones

Máster Universitario en Ingeniería Informática

Curso 2015 - 2016

Germán Martínez Maldonado

54097023B

germanm@correo.ugr.es

Objetivo N.º 1

Cálculo de estadísticos básicos

Utilizando como base el conjunto de datos ECBDL14 situado en la carpeta /user/daniel/ECBDL14_10columns/ECBDL14_10.data **obtener los siguiente valores:**

1. Calcula el valor mínimo de la variable (columna) 5

- Archivos: Min.java, MinMapper.java, MinReducer.java
- Salida:

Min: -13.0

2. Calcula el valor máximo de la variable (columna) 5

- Archivos: Max.java, MaxMapper.java, MaxReducer.java
- Salida:

Max: 10.0

3. Calcula al mismo tiempo los valores máximo y mínimo de la variable 5

- Archivos: MaxMin.java, MaxMinMapper.java, MaxMinReducer.java
- Salida:

Max: 10.0
Min: -13.0

4. Calcula los valores máximo y mínimo de todas las variables (salvo la última, que es la etiqueta de clase)

- Archivos: MaxMinAll.java, MaxMinAllMapper.java, MaxMinAllReducer.java
- Salida:

Max 1: 0.186
Min 1: 0.0
Max 2: 10.0
Min 2: -15.0
Max 3: 8.0
Min 3: -15.0
Max 4: 9.0
Min 4: -15.0
Max 5: 10.0
Min 5: -13.0
Max 6: 9.0
Min 6: -13.0
Max 7: 10.0
Min 7: -13.0
Max 8: 7.0
Min 8: -15.0
Max 9: 9.0
Min 9: -14.0
Max 0: 0.926
Min 0: 0.039

5. Realizar la media de la variable 5

- Archivos: Avg.java, AvgMapper.java, AvgReducer.java
- Salida:

Avg: -1.2354093582139625

6. Obtener la media de todas las variables (salvo la clase)

- Archivos: AvgAll.java AvgAllMapper.java AvgAllReducer.java
- Salida:

Avg 1: 0.05522961720265976
Avg 2: -2.146706610502992
Avg 3: -1.3587671160129455
Avg 4: -1.6588393726224624
Avg 5: -1.2354093582139625
Avg 6: -2.2260543199540925
Avg 7: -1.5609223677950506
Avg 8: -1.7251840805658227
Avg 9: -1.6793301243109373
Avg 0: 0.2508627955543515

7. Comprobar si el conjunto de datos EXBDL14 es balanceado o no balanceado, es decir, que el ratio entre las clases sea menor o mayor que 1.5 respectivamente.

- Archivos: Balanced.java, BalancedMapper.java, BalancedReducer.java
- Salida:

Ratio: 45.0
(El conjunto de datos no es balanceado dado que el ratio entre las clases es mayor que 1.5)

8. Cálculo del coeficiente de correlación entre todas las parejas de variables

Objetivo N.º 2

Minería de Datos en Big Data

Utilizando la herramienta Apache Mahout 1 , realizar los siguientes cálculos sobre el conjunto de datos KDDCUP'99 situado en vuestra carpeta raíz (dividido en dos partes: en 10% y 100%):

1. **Ejecutar el algoritmo "Random Forest" sobre ambos conjuntos de datos y comprueba el rendimiento alcanzado de acuerdo a los siguientes casos:**
   (a) Número de Maps:
   - 1 (secuencial), 32, 64

   (b) Número de árboles:
   - 100 árboles

2. **Del punto anterior, obtener una tabla que indique los siguientes datos:**

Ejecución	Características del modelo			Medidas de calidad				Tiempo de ejecución para entrenamiento
	Número de nodos	Número de nodos promedio	Profundidad máxima promedia	Kappa	Exactitud	Confiabilidad	Confiabilidad (desviación estándar)	
10%, 1 map	17759	177	7	-0.5406	99.9979%	66.6632%	0.5773	12m 9s 265
10%, 32 maps	6529	65	5	-0.5406	99.9804%	66.6559%	0.5773	16s 564
10% , 64 maps	4315	43	4	-0.5405	99.9279%	66.6341%	0.5771	18s 13
100%, 1 map	31818	318	9	-0.4721	99.9992%	66.6658%	0.5773	3h 16m 26s 999
100%, 32 maps	14442	144	7	-0.4721	99.9971%	66.6639%	0.5773	25s 107
100%, 64 maps	11082	110	6	-0.4721	99.9964%	66.6632%	0.5773	31s 766