Echidna Clustering

Frank Roger Salas Ticona

9 de noviembre de 2022

Clustering Mixed Data

Existen diversos algoritmos, para agrupar/clusterizar datos uniformes. Y es por ello que existe la necesidad de un algoritmo que nos permita clusterizar datos mixtos o no uniformes.
Un ejemplo de datos mixtos, es el *Tráfico en redes*.

Tenemos tipos de datos tales como: *SrcIP, DstIP, Protocol, SrcPort, DstPort, bytes*. Podemos observar, que existen tipos de datos jerárquicos como *SrcIP, DstIP*, *bytes* es cuantitativo y *Protocol, SrcPort, DstPort* son de tipo nominal y categóricos.

Distance Functions

1. **Atributos cuantitativos**: Son representados mediante escalares. Sus centroides en el cluster son dados por el promedio de N puntos. Calculamos la distancia entre clusteres usando la distancia Euclidiana.
2. **Atributos categóricos o cualitativos**: Un cluster con N puntos es representado por un histograma de frecuencias, de los atributos. De esta manera calculamos la distancia entre clusteres usando la distancia Euclidiana entre la frecuencia de cada atributo.
3. **Atributos jerárquicos**: Representado por un árbol. Cada nodo que no es una hoja es la generalización de nodos hoja en el subárbol en dicho nodo.

Distance Functions

En un cluster C cuyo atributo jerárquico corresponde a la dirección IP, es representado por un prefijo $I\bar{P}/p$. Calculamos la distancia enter dos clusteres con centroides, $I\bar{P}_1/p$ y $I\bar{P}_2/p$.

$$d_h(C_1, C_2) = 32 - p$$

Si $p > 8$ o 24 si $p \leq 8$, $p = CommonPrefix(I\bar{P}_1/p/I\bar{P}_2/p)$.

Radius Calculation

Para controlar la varianza de los datos en un cluster, necesitamos alguna medidad del *radio* del cluster. Podemos calcularlo para los atributos cuantitativos y cualitativos, con las desviación de los atributos en el cluster.
Para el caso de datos jerárquicos como los IPs, obtenemos *MinIP* y *MaxIP*. Obtenemos un rango tal que $C[i].range = (minIP, maxIP)$, medimos este *radio* con la altura del sub árbol más pequeño en la generalización jerárquica.

$$R_h = (32 - CommonPrefix(minIP, maxIP)/32)$$

.

Cluster formation

Echidna, es formado por un Árbol CF o CF-Tree [1].
Cada cluster como en BIRCH es representado por un vector que contiene estadísticas suficientes para calcular el *centroide* y *radio*. Cada dato X, comienza desde la raíz siguiendo el camino P hasta un nodo hoja. Se inserta al cluster más cercano y se actualiza en radio. En caso el número de entradas haya alcanzado el tope máximo entonces, el nodo se divide, y hacer el proceso recursivo en hasta llegar al nodo raíz.

Complexity

En un árbol balanceado por su altura, con un factor de *branching* B y m nodos, se harán $log_B m$ comparaciones que on requeridas al momento de realizar una insercion.

$$O(N * B(1 + log_B m))$$

Bibliografía

[1] Abdun Naser Mahmood, Christopher Leckie y Parampalli Udaya. "ECHIDNA: Efficient clustering of hierarchical data for network traffic analysis". En: *International Conference on Research in Networking*. Springer. 2006, págs. 1092-1098.