CLIQUE

Shuaiwu

Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, China
shuaiwu.ws@gmail.com

1 CLIQUE

CLIQUE (Clustering In QUEst)[2,1]is a density-based, grid-based clustering algorithm, which employs bottom-up subspace search method to find clusters in subspaces of the original dataset. It's the bottom-up strategy which can effectively reduce the search subspace[3]. Monotonicity is key principle of bottom-up search, which is showed in Theorem 1.

Theorem 1. *A set of points is a cluster in k-D space must be a part of a cluster in any (k-1)-D projection of the space.*

The algorithm begins with determining 1-D cells with densities above a fixed threshold τ . After finding (k-1)-D dense units, the k-D candidate units will be generated through the candidate process given in **Algorithm 2** and the non-dense units will be deleted from the candidate set. The algorithm will terminate when the candidate set is empty.

To lower the time complexity, MDL(Minimal Description Length) principle is introduced to prune some subspaces with a relatively low coverage. The coverage of subspace S_i is defined as $x_{S_i} = \sum_{u_i \in S_i} count(u_i)$, where u_i is dense unit of S_i and $count(u_i)$ is the number of points in the unit u_i. All subspace in k-D will be sorted in the descending order of their coverage. The sorted subspaces need to be divided into categories: the selected set I and the pruned set P. But how to get a good cutting point is extremely important. If we are required to store the coverage information of the two sets I and P. The mean of each set and the difference between each subspace and the mean of belonged set need to be included in the code. The length of code CL is the sum of the bit lengths of stored numbers. Assume we decide to prune subspace S_{i+1},$\cdots$,S_n, the encoding length is:

$$CL(i) = CL_I(i) + CL_P(i) \tag{1}$$

where

$$CL_I(i) = log_2(\mu_I(i)) + \sum_{1 \leq j \leq i} log_2(\left|x_{S_j} - \mu_I(i)\right|) \tag{2}$$

$$CL_P(i) = log_2(\mu_P(i)) + \sum_{i+1 \leq j \leq n} log_2(\left|x_{S_j} - \mu_P(i)\right|) \tag{3}$$

$$\mu_I(i) = \frac{\sum_{1 \leq j \leq i} x_{S_j}}{i} \tag{4}$$

$$\mu_P(i) = \frac{\sum_{i+1 \leq j \leq n} x_{S_j}}{n - i} \tag{5}$$

The optimal cutting point i will have the minimal encoding length.

After finding all dense units, a depth-first search algorithm is used to find clusters in each subspace. To make the representation of clusters more interpretable, CLIQUE generate minimal description for each cluster. A minimal description of a cluster is a non-redundant covering of the cluster with maximal regions. A set of k-D regions R is a cover of k-D cluster C if every region of R is included in C and each units of C is contained in at least one of the regions of R. CLIQUE greedily covers a cluster by a number of maximal hyper-rectangles, and then deletes redundant rectangles to get minimal description.

The algorithm is listed below:

Algorithm 1 CLIQUE (CLustering In QUEst)

```
Input: $Data$: a dataset $S$, $\xi$:the number units for each dimension, $\tau$:the density threshold
Determine 1-D Dense units set $C_1$ and copy $C_1$ to $D_1$
k = 2
While $D_{k-1}$ is not null
    Generate the candidate units set $C_k$ for k-D with Algorithm 2
    Discard the non-dense units in $C_k$ and copy $C_k$ to $D_k$
    Discard the units of $D_k$ in the subspaces which are determined to discard according to MDL
    k = k+1
END
FOR each subspace
    Find clusters in the subspace with Algorithm 3
    Generate a minimal description for each cluster of the subspace
END
Output: the clustering result for each subspace
```

Algorithm 2 Candidate generator

```
Input: $D_{k-1}$: the dense $(k-1)$-$D$ units
From $D_{k-1}$ select two units $u_1$ and $u_2$ where $u_1.a_1 = u_2.a_1$, $u_1.l_1 = u_2.l_1$, $u_1.h_1 = u_2.h_1, \cdots, u_1.a_{k-2} = u_2.a_{k-2}$, $u_1.l_{k-2} = u_2.l_{k-2}$, $u_1.h_{k-2} = u_2.h_{k-2}$, $u_1.a_{k-1} < u_2.a_{k-1}$
    % u. $[l_i, h_i)$ represents the interval in the $i^{th}$ dimension of unit $u$ and the relation $<$ represents lexicographic ordering on attributes
Copy $u_1$ to $u_{new}$ and add a new dimenion for $u_{new}$, $u_{new}.a_k = u_2.a_{k-1}$, $u_{new}.l_k = u_2.l_{k-1}$, $u_{new}.h_k = u_2.h_{k-1}$
Output: $C_k$: the set of candidate dense units for $k$-$D$
```

Algorithm 3 Identication of clusters

Input: D: a set of dense units in k-dimensional space S,

```
Set all u.label to null and index to 1
FOR i = 1:N
    IF $u_i.label$ is null
        dfs($u_i$,index)
    END
index = index + 1
END
FOR j = 1:index-1
    insert all units into a cluster with u.label = j into $D^j$
END
```

Output: a partition of D into $D^1, \cdots, D^q$, such all units in D^i are connected and no units $u^i \in D^i$, $u^j \in D^j$ with $i \neq j$ are connected.

Algorithm 4 Depth-first search (dfs)

Input: u: a units in k-dimensional space S (u =$\{[l_1, h_1); \cdots ; [l_k, h_k)\}$, index: the index of clusters

```
u.num = index
FOR i = 1:k
% $u^l$ = { $[l_1, h_1); \cdots ; [l^l_k, h^l_k); \cdots ; [l_k, h_k)$ }
    IF ($u^l$ is dense ) and ($u^l.label$ is null)
        dfs($u^l$,index)
    END
% $u^r$ = { $[l_1, h_1); \cdots ; [l^r_k, h^r_k); \cdots ; [l_k, h_k)$ }
    IF ($u^r$ is dense ) and ($u^r.label$ is null)
        dfs($u^r$,index)
    END
END
```

Output: null

References

1. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. pp. 94–105. SIGMOD '98, ACM, New York, NY, USA (1998). https://doi.org/10.1145/276304.276314, http://doi.acm.org/10.1145/276304.276314
2. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data. Data Mining and Knowledge Discovery **11**(1), 5–33 (Jul 2005). https://doi.org/10.1007/s10618-005-1396-1, https://doi.org/10.1007/s10618-005-1396-1
3. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: A review. SIGKDD Explor. Newsl. **6**(1), 90–105 (Jun 2004). https://doi.org/10.1145/1007730.1007731, http://doi.acm.org/10.1145/1007730.1007731