Automatic Subspace Clustering of High Dimensional Data for Data Mining Applicatyions

Li Cheng

Background

- The Curse Of Dimensionality
- Some solutions
  - Data Projection, Dimension Reduction, signature encoding
    - PCA, Wavelet, NN (SOM)
  - Feature Selection
- CLIQUE need not do that

The Contribution of CLIQUE

- Automatically find ***subspaces*** with high-density clusters in high dimensional attribute spaces

Some Definitions:

- A cluster is a maximal set of connected dense units in K-dimensions.
- Two K-dimensional units u_1, u_2 are connected if they have a common face, or if there exists other K-dim unit u_i, such that u_1, u_i and u_2 are connected consequently.
- A region in K dimensions is an axis-parallel rectangular K-dimensional set.

What is CLIQUE

- The basic idea is similar to APRIORI, the association rule algorithm.
  - A bottom-up scheme.
  - The Monotonicity Lemma
  - Prune to eliminate some outlines that their "support" is too small. The threshold here called "optimal cut point i"

Next

Flow Chart of CLIQUE

- Bottom-up to find dense units
- Further Prune subspaces using MDL principle
- Generating Minimal number of Regions, each region cover one cluster
  - Firstly, greedily find a number of maximal rectangles
  - Generate a minimal cover

Apriori algorithm

- Transaction Data :

  {1,4,5},{1,2},{3,4,5},{1,2,4,5}

  $L_1 = \{\{1\},\{2\},\{3\},\{4\}\}$

  $\Downarrow$ Cartesian Product

  $C_2 = \{\{1,2\},\{1,4\},\{1,5\},\{2,4\},\{2,5\},\{4,5\}\}$

  $\Downarrow$ Support Counting

  $L_2 = \{\{1,2\},\{1,4\},\{1,5\},\{4,5\}\}$

  $\Downarrow$ Join

  $(\{\{1,2,4\},\{1,2,5\},\{1,4,5\}\})$

  $\Downarrow$ Not Large $\Rightarrow$ Pruning

  $C_3 = \{\{1,4,5\}\}$

  $\Downarrow$ Support Counting

  $L_3 = \{\{1,4,5\}\}$

- Reproduced from http://www.scs.carleton.ca/~kimasaki/DataMining/summary/

Basic Idea of CLIQUE

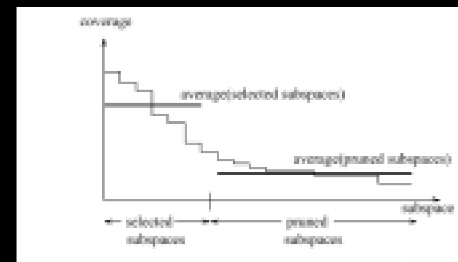
Monotonicity:

If a collection of points S is a cluster in a K-dimensional space, then S is also part of a cluster in any (k-1) dimensional projections of this space.

Flow Chart of CLIQUE

- Bottom-up to find dense units
- Further Prune subspaces using MDL principle
- Generating Minimal number of Regions, each region cover one cluster
  - Firstly, greedily find a number of maximal rectangles
  - Generate a minimal cover

Prune subspaces using MDL principle



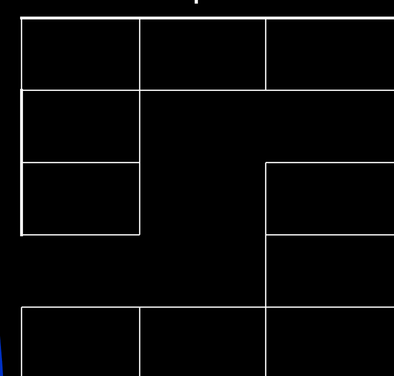
- Partitioning of the subspaces into selected and prune sets

Flow Chart of CLIQUE

- Bottom-up to find dense units
- Further Prune subspaces using MDL principle
- Generating Minimal number of Regions, each region cover one cluster
  - Firstly, greedily find a number of maximal rectangles
  - Generate a minimal cover

Flow Chart of CLIQUE (Cont.)

- An Example:



Pro. And Con.

- Pro.
  - Order Insensitive
  - Arbitrary Shape of Clusters
  - Tolerant of missing values
  - Doesn't presume some canonical distribution
  - Scalability O(n)
  - Insensitive to noise

Pro. And Con. (Cont.)

- Cons.
  - Some parameters that hard to pre-select: ξ (partition threshold) and τ (density threshold, i.e. support threshold)
  - Prone to higher dimensional clusters
  - Some potential clusters will lost in the density-units or subspace-prune procedures

Comparison with Birch, DBScan and PCA (SVD)

Concludes that CLIQUE performs better than Birch, DBScan and SVD

Table 1: BIRCH experimental results.

Dim. of data	Dim. of clusters	No. of clusters	Clusters found	True clusters identified
5	5	5	5	5
10	5	5	5	5
20	5	5	3,4,5	0
30	5	5	3,4	0
40	5	5	3,4	0
50	5	5	3	0

Table 2: DBSCAN experimental results.

Dim. of data	Dim. of clusters	No. of clusters	Clusters found	True clusters identified
5	5	5	5	5
7	5	5	5	5
8	5	5	3	1
10	5	5	1	0

Table 3: SVD decomposition experimental results.

Dim. of data (d)	Dim. of clusters	No. of clusters	$r_{d/2}$	$r_{(d-5)}$	$r_{(d-1)}$
10	5	5	0.647	0.647	0.907
20	5	5	0.606	0.827	0.969
30	5	5	0.583	0.858	0.972
40	5	5	0.557	0.897	0.981
50	5	5	0.552	0.919	0.984