CS685: Data Mining
Grid-Based Clustering Methods

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs685/

1st semester, 2020-21

Mon 1030-1200 (online)

Grid-Based Clustering Methods

- Space is partitioned into grids or cells
- Generally, cells are of uniform size
- Number of cells is independent of data distribution

Grid-Based Clustering Methods

- Space is partitioned into grids or cells
- Generally, cells are of uniform size
- Number of cells is independent of data distribution
- Clusters are formed by examining neighboring grids
- Results in faster methods that depend less on the volume of data but more on the number of cells

Grid-Based Clustering Methods

- Space is partitioned into grids or cells
- Generally, cells are of uniform size
- Number of cells is independent of data distribution
- Clusters are formed by examining neighboring grids
- Results in faster methods that depend less on the volume of data but more on the number of cells
- Grids may be at multiple resolutions

Grid-Based Clustering Methods

- Space is partitioned into grids or cells
- Generally, cells are of uniform size
- Number of cells is independent of data distribution
- Clusters are formed by examining neighboring grids
- Results in faster methods that depend less on the volume of data but more on the number of cells
- Grids may be at multiple resolutions
- Statistical Information Grid (STING)
- Clustering in Quest (CLIQUE)

STING

- STatistical INformation Grid (STING)
- Grids at multiple resolutions
- Forms a hierarchy
  - Cells in higher level are partitioned to form cells in lower levels

STING

- STatistical INformation Grid (STING)
- Grids at multiple resolutions
- Forms a hierarchy
  - Cells in higher level are partitioned to form cells in lower levels
- Each cell maintains certain *statistical* parameters
- Attribute-independent
  - Count

STING

- STatistical INformation Grid (STING)
- Grids at multiple resolutions
- Forms a hierarchy
  - Cells in higher level are partitioned to form cells in lower levels
- Each cell maintains certain *statistical* parameters
- Attribute-independent
  - Count
- Attribute-dependent
  - Mean
  - Standard deviation
  - Minimum
  - Maximum
  - Distribution (*none* if not known)

STING

- STatistical INformation Grid (STING)
- Grids at multiple resolutions
- Forms a hierarchy
  - Cells in higher level are partitioned to form cells in lower levels
- Each cell maintains certain *statistical* parameters
- Attribute-independent
  - Count
- Attribute-dependent
  - Mean
  - Standard deviation
  - Minimum
  - Maximum
  - Distribution (*none* if not known)
- Parameters at higher level cells can be computed from lower levels

Clustering

- STING is a generalized framework for answering spatial queries
- A spatial query with constraints is first applied on a particular level
- For each cell in that level, a *confidence interval* is computed to ascertain its relevance to the query
- Irrelevant cells are removed
- For relevant cells, the query is drilled down to lower levels till the bottom-most layer is reached
- Cells satisfying the query are returned

Clustering

- STING is a generalized framework for answering spatial queries
- A spatial query with constraints is first applied on a particular level
- For each cell in that level, a *confidence interval* is computed to ascertain its relevance to the query
- Irrelevant cells are removed
- For relevant cells, the query is drilled down to lower levels till the bottom-most layer is reached
- Cells satisfying the query are returned
- How is clustering done?

Clustering

- STING is a generalized framework for answering spatial queries
- A spatial query with constraints is first applied on a particular level
- For each cell in that level, a *confidence interval* is computed to ascertain its relevance to the query
- Irrelevant cells are removed
- For relevant cells, the query is drilled down to lower levels till the bottom-most layer is reached
- Cells satisfying the query are returned
- How is clustering done?
- Clustering can be viewed as a spatial query with constraints
  - Diameter of a cluster should be less than δ
  - Density of a cluster should be greater than τ

Discussion

- Very fast

Discussion

- Very fast
- Computing of cell parameters at bottom levels require only one pass on the data, i.e., $O(n)$

Discussion

- Very fast
- Computing of cell parameters at bottom levels require only one pass on the data, i.e., $O(n)$
- Processing grids is $O(g)$ where g is the total number of cells
- Therefore, total time is $O(n + g)$

Discussion

- Very fast
- Computing of cell parameters at bottom levels require only one pass on the data, i.e., $O(n)$
- Processing grids is $O(g)$ where g is the total number of cells
- Therefore, total time is $O(n + g)$
- Depends on resolution of grids, especially the bottom most layer

Discussion

- Very fast
- Computing of cell parameters at bottom levels require only one pass on the data, i.e., $O(n)$
- Processing grids is $O(g)$ where g is the total number of cells
- Therefore, total time is $O(n + g)$
- Depends on resolution of grids, especially the bottom most layer
- For high dimensions, number of cells may be too high

Discussion

- Very fast
- Computing of cell parameters at bottom levels require only one pass on the data, i.e., $O(n)$
- Processing grids is $O(g)$ where g is the total number of cells
- Therefore, total time is $O(n + g)$
- Depends on resolution of grids, especially the bottom most layer
- For high dimensions, number of cells may be too high
- Clusters returned are isothetic, i.e., they are always aligned across axes

Discussion

- Very fast
- Computing of cell parameters at bottom levels require only one pass on the data, i.e., $O(n)$
- Processing grids is $O(g)$ where g is the total number of cells
- Therefore, total time is $O(n + g)$
- Depends on resolution of grids, especially the bottom most layer
- For high dimensions, number of cells may be too high
- Clusters returned are isothetic, i.e., they are always aligned across axes
- Can identify non-convex clusters

CLIQUE

- CLustering In QUEst (CLIQUE)
- Uses *Apriori* property: If a set of points forms a density-based cluster in k dimensions, they must form density-based clusters in *every* subset of dimensions
- This is the *monotonicity* property of subspace clustering

CLIQUE

- CLustering In QUEst (CLIQUE)
- Uses *Apriori* property: If a set of points forms a density-based cluster in k dimensions, they must form density-based clusters in *every* subset of dimensions
- This is the *monotonicity* property of subspace clustering
- A cell in k dimensions can have $\geq \tau$ points if and only if every possible $(k-1)$-dimensional cell projected from it has $\geq \tau$ points
- Apriori-like algorithm

CLIQUE

- CLustering In QUEst (CLIQUE)
- Uses *Apriori* property: If a set of points forms a density-based cluster in k dimensions, they must form density-based clusters in *every* subset of dimensions
- This is the *monotonicity* property of subspace clustering
- A cell in k dimensions can have $\geq \tau$ points if and only if every possible $(k-1)$-dimensional cell projected from it has $\geq \tau$ points
- Apriori-like algorithm
- Starts by identifying *dense* intervals in $k = 1$ dimension
- Generate all possible $(k+1)$-dimensional cells
- Prune cells that fail the density criterion
- Continue till k is exhausted or no dense cells

Discussion

- Generates all clusters in every subspace

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality
- Initial length of interval in each dimension is an important parameter

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality
- Initial length of interval in each dimension is an important parameter
- So is the density threshold

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality
- Initial length of interval in each dimension is an important parameter
- So is the density threshold
- Can potentially take exponential time

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality
- Initial length of interval in each dimension is an important parameter
- So is the density threshold
- Can potentially take exponential time
- Works better for sparse datasets

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality
- Initial length of interval in each dimension is an important parameter
- So is the density threshold
- Can potentially take exponential time
- Works better for sparse datasets
- Clusters in same space can overlap

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality
- Initial length of interval in each dimension is an important parameter
- So is the density threshold
- Can potentially take exponential time
- Works better for sparse datasets
- Clusters in same space can overlap
- May use *maximal* clusters to merge

Discussion

- Generates all clusters in every subspace
- Density criterion, i.e., number of points in a cell remains constant across dimensionality
- Initial length of interval in each dimension is an important parameter
- So is the density threshold
- Can potentially take exponential time
- Works better for sparse datasets
- Clusters in same space can overlap
- May use *maximal* clusters to merge
- Instead of density, may work with entropy, etc.