Drawback of Distance-based Methods

- Hard to find clusters with irregular shapes
- Hard to specify the number of clusters
- Heuristic: a cluster must be dense
**Directly Density Reachable**

- Parameters
  - $\text{Eps}$: Maximum radius of the neighborhood
  - $\text{MinPts}$: Minimum number of points in an Eps-neighborhood of that point
  - $\text{NEps}(p) = \{ q \mid \text{dist}(p, q) \leq \text{Eps} \}$
  - Core object $p$: $|\text{NEps}(p)| \geq \text{MinPts}$
  - Point $q$ directly density-reachable from $p$ iff $q \in \text{NEps}(p)$ and $p$ is a core object

**Density-Based Clustering: Background (II)**

- Density-reachable
  - Directly density reachable $p_1 \rightarrow p_2, p_2 \rightarrow p_3, \ldots, p_{n-1} \rightarrow p_n \Rightarrow p_n$ density-reachable from $p_1$

- Density-connected
  - Points $p, q$ are density-reachable from $o \Rightarrow p$ and $q$ are density-connected
DBSCAN

- A cluster: a maximal set of density-connected points
  - Discover clusters of arbitrary shape in spatial databases with noise

- Arbitrary select a point \( p \)
- Retrieve all points density-reachable from \( p \) wrt \( \varepsilon \) and \( \text{MinPts} \)
- If \( p \) is a core point, a cluster is formed
- If \( p \) is a border point, no points are density-reachable from \( p \) and DBSCAN visits the next point of the database
- Continue the process until all of the points have been processed

\( \text{Eps} = 1 \text{cm} \)
\( \text{MinPts} = 5 \)
Problems of DBSCAN

- Different clusters may have very different densities
- Clusters may be in hierarchies

OPTICS: A Cluster-ordering Method

- OPTICS: ordering points to identify the clustering structure
- “Group” points by density connectivity
  - Hierarchies of clusters
- Visualize clusters and the hierarchy
DENCLUE: Using Density Functions

- DENsity-based CLUstEring
- Major features
  - Solid mathematical foundation
  - Good for data sets with large amounts of noise
  - Allow a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
  - Significantly faster than existing algorithms (faster than DBSCAN by a factor of up to 45)
  - But need a large number of parameters

Grid-based Clustering Methods

- Ideas
  - Using multi-resolution grid data structures
  - Use dense grid cells to form clusters
- Several interesting methods
  - STING
  - WaveCluster
  - CLIQUE
STING: A Statistical Information Grid Approach

The spatial area area is divided into rectangular cells.
There are several levels of cells corresponding to different levels of resolution.

STING: A Statistical Information Grid Approach (2)

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level.
- Statistical information of each cell is calculated and stored beforehand and is used to answer queries.
- Parameters of higher level cells can be easily calculated from parameters of lower level cell:
  - count, mean, s, min, max
  - type of distribution—normal, uniform, etc.
- Use a top-down approach to answer spatial data queries.
- Start from a pre-selected layer—typically with a small number of cells.
- For each cell in the current level compute the confidence interval.
STING: A Statistical Information Grid Approach (3)

- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached

Advantages:
- Query-independent, easy to parallelize, incremental update
- $O(K)$, where $K$ is the number of grid cells at the lowest level

Disadvantages:
- All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected
WaveCluster

- A multi-resolution clustering approach which applies wavelet transform to the feature space
  - A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-bands.
- Both grid-based and density-based
- Input parameters:
  - # of grid cells for each dimension
  - the wavelet, and the # of applications of wavelet transform.

How to apply wavelet transform to find clusters

- Summaries the data by imposing a multidimensional grid structure onto data space
- These multidimensional spatial data objects are represented in an n-dimensional feature space
- Apply wavelet transform on feature space to find the dense regions in the feature space
- Apply wavelet transform multiple times which result in clusters at different scales from fine to coarse
Wavelet Transform

- Decomposes a signal into different frequency subbands. (can be applied to n-dimensional signals)
- Data are transformed to preserve relative distance between objects at different levels of resolution.
- Allows natural clusters to become more distinguishable

What Is Wavelet (2)?

![Diagram of wavelet transform with frequency and time axes]
Quantization

Figure 1: A sample 2-dimensional feature space.

Transformation

COMP 790-090 Data Mining: Concepts, Algorithms, and Applications
WaveCluster

- Why is wavelet transformation useful for clustering
  - Unsupervised clustering
    It uses hat-shape filters to emphasize region where points cluster, but simultaneously to suppress weaker information in their boundary

Cohen-Daubechies-Feauveau (2,2) biorthogonal wavelet.

WaveCluster

- Effective removal of outliers
WaveCluster

- Multi-resolution

![Block diagram of multi-resolution wavelet transform.]

- Cost efficiency

WaveCluster
WaveCluster

- Major features:
  - Complexity $O(N)$
  - Detect arbitrary shaped clusters at different scales
  - Not sensitive to noise, not sensitive to input order
  - Only applicable to low dimensional data

CLIQUE (Clustering In QUEst)

- Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space
- CLIQUE can be considered as both density-based and grid-based
  - It partitions each dimension into the same number of equal length interval
  - It partitions an $m$-dimensional data space into non-overlapping rectangular units
  - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
  - A cluster is a maximal set of connected dense units within a subspace
CLIQUE: The Major Steps

- Partition the data space and find the number of points that lie inside each cell of the partition.
- Identify the subspaces that contain clusters using the Apriori principle.
- Identify clusters:
  - Determine dense units in all subspaces of interests
  - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
  - Determine maximal regions that cover a cluster of connected dense units for each cluster
  - Determination of minimal cover for each cluster
Strength and Weakness of CLIQUE

**Strength**
- It automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces.
- It is insensitive to the order of records in input and does not presume some canonical data distribution.
- It scales linearly with the size of input and has good scalability as the number of dimensions in the data increases.

**Weakness**
- The accuracy of the clustering result may be degraded at the expense of simplicity of the method.
Constrained Clustering

- Constraints exist in data space or in user queries
- Example: ATM allocation with bridges and highways
  - People can cross a highway by a bridge

Clustering With Obstacle Objects

*Not* Taking obstacles into account  Taking obstacles into account
Outlier Analysis

- “One person’s noise is another person’s signal”
- Outliers: the objects considerably dissimilar from the remainder of the data
  - Examples: credit card fraud, Michael Jordon, etc
  - Applications: credit card fraud detection, telecom fraud detection, customer segmentation, medical analysis, etc

Statistical Outlier Analysis

- Discordancy/outlier tests
  - 100+ tests proposed
- Data distribution
  - Distribution parameters
- The number of outliers
- The types of expected outliers
  - Example: upper or lower outliers in an ordered sample
### Drawbacks of Statistical Approaches

- Most tests are univariate
  - Unsuitable for multidimensional datasets
- All are distribution-based
  - Unknown distributions in many applications

![Probability Distribution Diagram](image)

### Depth-based Methods

- Organize data objects in layers with various depths
  - The shallow layers are more likely to contain outliers
- Example: Peeling, Depth contours
- Complexity $O(N^{\lceil k/2 \rceil})$ for k-d datasets
  - Unacceptable for $k>2$
Distance-based Outliers

- A DB(p, D)-outlier is an object O in a dataset T s.t. at least fraction p of the objects in T lies at a distance greater than distance D from O.
- Algorithms for mining distance-based outliers
  - The index-based algorithm, the nested-loop algorithm, the cell-based algorithm

Index-based Algorithms

- Find DB(p, D) outliers in T with n objects
  - Find an objects having at most $\lceil n(1-p) \rceil$ neighbors with radius D
- Algorithm
  - Build a standard multidimensional index
  - Search every object O with radius D
    - If there are at least $\lceil n(1-p) \rceil$ neighbors, O is not an outlier
    - Else, output O
Pros and Cons of Index-based Algorithms

- Complexity of search $O(kN^2)$
  - More scalable with dimensionality than depth-based approaches
- Building a right index is very costly
  - Index building cost renders the index-based algorithms non-competitive

A Naïve Nested-loop Algorithm

- For $j=1$ to $n$ do
  - Set $count_j = 0$;
  - For $k=1$ to $n$ do if $(dist(j,k) < D)$ then $count_j++$;
  - If $count_j \leq \lfloor n(1-p) \rfloor$ then output $j$ as an outlier;
- No explicit index construction
  - $O(N^2)$
- Many database scans
Optimizations of Nested-loop Algorithm

- Once an object has at least $\lfloor n(1-p) \rfloor$ neighbors with radius D, no need to count further.
- Use the data in main memory as much as possible.
  - Reduce the number of database scans.

References (1)

References (2)

- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB’94.
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB’98.
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. SIGMOD’96.