Clustering in Context

Kari Laasonen

Kari.Laasonen@cs.helsinki.fi

University of Helsinki – Department of Computer Science

General Approaches

1. Detect typical contexts
   - Clustering
   - Goal: help users to adapt to context changes
   - Presentable information
2. Predict appropriate behavior
   - Classification/machine learning
   - Goal: the device reacts to context changes automatically
   - Recommendations for the user (acting without user consent is risky)

Clustering Context Data

- Online clustering
  - Clusters are updated with each new data point
  - This could be done with BIRCH-like summaries
  - Initially: collect data for time T, then analyze the results
- Amount of data
  - Phone memory is a scarce resource
  - But does a single user generate enough data?
- Categorical data and metrics

Projective Clustering

Projective Clustering

- Assume data points are d-dimensional
- Geometric intuition breaks down for large d
- "Curse of dimensionality:" everything is far away
- Projective (or subspace) clustering attempts to find relevant dimensions or subspaces
- Project the data on a D-dimensional subspace ($D \ll d$) and see what it looks like
- Subspaces are typically aligned with coordinate axes; this is easier and works well in practice.

CLIQUE

Devised by Agrawal et al., ACM SIGMOD '98

- Clusters have a higher density of points than its surroundings
- To detect areas of high density, partition the space into cells
  - On the first round, cells are one-dimensional
  - Dense cells of dimension $k-1$ are combined to generate candidate cells of dimension k.
  - A database pass prunes sparse cells
- Strongly resembles the Apriori algorithm for finding frequent sets

Monte Carlo Approach

Procopiuc et al., ACM SIGMOD '02

- Greedy: finds one optimal cluster.
- Generates candidate clusters randomly:
  - Choose a point p in the cluster
  - Choose a discriminating subset X
  - If dimension i is in the cluster, then $|q_i - p_i| \leq w$ for all $q \in X$
  - Otherwise there is $q \in X$ such that $|q_i - p_i| > w$.
- Choose the candidate that maximizes a goodness metric.

Monte Carlo Approach

- Repeat the process to boost success probability.
- For typical parameter values, the number of candidates is $O(d^5)$ and $|X| = O(\log d)$.
- Other approaches are worse: the running time of (plain) CLIQUE is $O(c^k)$, where k is the highest dimensionality of dense cells.

Issues to Consider

- Parameters
  - Density threshold
  - Cell size/box width
  - Balance between number of points and number of dimensions
- Efficiency