Error-Aware Density-Based Clustering of Imprecise Measurement Values

Dirk Habich, Peter B. Volk, Wolfgang Lehner
Dresden University of Technology
Database Technology Group
Dresden, Germany

Ralf Dittmann, Clemens Utzny
Advanced Mask Technology Center
Dresden, Germany
Outline

• Introduction
  – Application Area
  – Problem Specification including Related Work

• Error-Aware Clustering
  – DBSCAN Extension
  – Quality Measures
  – Two-Phase Clustering Approach

• Evaluation
• Summary and Outlook
• **Advanced Mask Technology Center (AMTC)**
  - leading center for research and development of new generation of photolithographic masks
  - no mass production, but specific customer-oriented production

  - overall aim
    • produce exactly one mask per order that can be shipped to the customer
  - overall problem
    • determination of manufacturing parameters
- Production control based on historical masks
Specific Task

- **Task**
  - masks with similar behavior according to manufacturing process should be found
  → Data Clustering

- **Problem Specification**
  - data objects are described by n-dimensional regions (figure b), not by n-dimensional points (figure a)
  - data regions may have arbitrary shapes
Evaluation of Impact of Uncertainty

- **Evaluation based**
  - on synthetic generated two-dimensional data sets and
  - well-defined uncertainty for each dimension

  - derived data sets:
    - changing the position of data points within the data region specified by the uncertainty
    - include only points that are moved by a certain percentage of the allowed maximum

  - clustering with DBSCAN, whereas parameters of are determined using the proposed heuristic
• **Results**
  - compared clustering results by measuring the disagreement
  - distance function (Gionis et al. ICDE 2005)
    • $C_1$, $C_2$ clusters
    • $u$, $v$ points

\[
  d_{u,v}(C_1, C_2) = \begin{cases} 
    1 & \text{if } C_1(u) = C_1(v) \text{ and } C_2(u) \neq C_2(v), \\
    & \text{or } C_1(u) \neq C_1(v) \text{ and } C_2(u) = C_2(v), \\
    0 & \text{otherwise.} 
  \end{cases}
\]

\[
  d_V(C_1, C_2) = \sum_{(u,v) \in V \times V} d_{u,v}(C_1, C_2).
\]
• **Results**
  - Experiment 1: four satellites are very close to the center cluster
  - Experiment 2: four satellites are more remote from the center cluster
Related Work

- **Clustering Uncertain Data**
  - partition-based: UKMeans
  - density-based: fDBSCAN
  - Hierarchical density-based: fOptics

- **Properties**
  - objects are represented by pdf’s
  - distance function: expected distance metric (high complexity)
Outline

• **Introduction**
  – Application Area
  – Problem Specification including Related Work

• **Error-Aware Clustering**
  – DBSCAN Extension (DBSCAN$^{EA}$)
  – Quality Measures
  – Two-Phase Clustering Approach

• **Evaluation**
• **Summary and Outlook**
- **Density-Based Clustering Approach**
  - Based on DBSCAN

- **Similarity Measure between two Uncertain Objects**
  - Approximation of Objects using Minimum Bounding Rectangle (MBR)
  - Minimum and Maximum Distance
    - based on MBR
Distance Function

- **Two Extremes**
  - maximum distance = pessimistic value
  - minimum distance = optimistic value

- **Distance Function**
  - combination of maximum and minimum values
    \[ dist(P, Q) = (1 - \lambda) \times d_{\text{max}}(P, Q) + \lambda \times d_{\text{min}}(P, Q) \]

- **Correlation Factor \( \lambda \)**
  - domain: \( 0 \leq \lambda \leq 1 \)
  - \( \lambda = 0 \rightarrow \) pessimistic
  - \( \lambda = 1 \rightarrow \) optimistic
• applying regular DBSCAN with specific value of $\lambda$
  – slightly adapted definition of (DBSCAN\textsuperscript{EA})
    • $(\varepsilon, \lambda)$-neighborhood of a data region
    • directly density-reachable
    • density-reachable
    • density-connected
    • density-based cluster region

• Parameter of DBSCAN\textsuperscript{EA}
  – $\varepsilon$, MinPts, $\lambda$
Example

- clustering results with different values for
  - synthetic generated two-dimensional data set
  - symmetric errors for all dimensions (hybercube)

- Conclusion: different values of produce different clustering results
- Problem: DBSCAN$^{EA}$ considers only one possible value of
Quality Measures

- **Parameter of DBSCAN**
  - $\varepsilon$, MinPts, $\lambda$

- **Variations of $\lambda$ allows to derive further means of quality measures**
  - Cluster Stability
  - Similarity Measure

- **Cluster Stability – CS**
  - range of $\lambda$ in which cluster is stable
  - calculation
    - applying DBSCAN$^{EA}$ multiple times on the data sets
    - increasing lambda by $\Delta\lambda$
    - comparing clustering results
• **Similarity Measure**
  - determines how close two data uncertain objects are associated with each other among different values of $\lambda$
  - calculation
    • applying DBSCAN$^{EA}$ multiple times on the data sets
    • decreasing $\lambda$ from $\lambda=1$ to $\lambda=0$ by $\Delta\lambda$
    • for each clustering result
      - Compute a local similarity
      - $P$ and $Q$ in the same cluster
        $$\text{LS}(P,Q) = (\Delta\lambda)^{(1-\lambda)}$$
      - $P$ and $Q$ in different clusters
        $$\text{LS}(P,Q) = 0$$
    • Global similarity
      - $\text{GS}(P,Q) = \text{SUM}($$\text{LS}(P,Q))$$
Extended Error-Aware Clustering

- **Two-Phase Clustering**
  - first phase: number of local clustering using DBSCAN\textsuperscript{EA}
  - second phase:
    - constructing similarity graph
    - Computing global clustering result
      - Clustering Aggregation
      - Gionis et al. ICDE 2005
      - Habich et al. SAC 2006
      - ...
Outline

- **Introduction**
  - Application Area
  - Problem Specification including Related Work

- **Error-Aware Clustering**
  - DBSCAN Extension
  - Quality Measures
  - Two-Phase Clustering Approach

- **Evaluation**
- **Summary and Outlook**
Evaluation

- **AMTC**
  - implemented
  - parameter prediction based on clustering result shows improvement

- **Synthetic Data Set**
Conclusion

• **Presented**
  - Application area (Advanced Mask Technology Center)
  - Related Work
  - DBSCAN<sup>EA</sup>
  - Two-Phase Clustering Approach

• **TODO**
  - More evaluation
  - Refinement of approach