The Generalized MDL Approach for Summarization

Laks V.S. Lakshmanan (UBC)
Raymond T. Ng (UBC)
Christine X. Wang (UBC)
Xiaodong Zhou (UBC)
Theodore J. Johnson (AT&T Research)

(Work supported by NSERC and NCE/IRIS.)
Overview

• Introduction
• Motivation & Problem Statement
• Spatial Case – MDL & GMDL
• Experiments ← X
• Categorical Case
• More Experiments ← X
• Related work
• Summary and Related/Future Work
Introduction

• How best to convey large answer sets for queries?
  – Simple enumeration: accurate but not necessarily most useful
  – Summaries: not (necessarily) 100% accurate but can be more intuitive

• Why is this problem interesting?
  – OLAP queries over multi-dimensional data typically produce data intensive answers
• **Example**: (i) customer segmentation based on buying pattern

- too many answers, in general
- solution: summarize
- description via range constraints
  \[ \Rightarrow \text{axis-parallel hyper-rectangles} \]
  \[ \Rightarrow \text{most concise = MDL} \]
• **Example:** (ii) aggregate sales performance analysis

- ≥ 2 * last year’s sales
- description via hierarchical ranges = tuples of nodes
- most concise = MDL
Motivation

- **Examples**: (i) customer segmentation based on buying pattern

```
<table>
<thead>
<tr>
<th>X</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>
```

- frequency ≥ t
- X frequency < t/2
- “white” otherwise
- white budget = 2
- white budget ≥ 10
• **Example:** (ii) aggregate sales performance analysis

- ≥ 2 * last year’s sales
- description via hierarchical ranges = tuples of nodes
- most concise = MDL
**Motivation (contd.)**

- **Example:** (ii) aggregate sales performance analysis

  - $\geq 2 \times$ last year’s sales
  - $X < \frac{1}{2} \times$ last year’s sales

  ![Diagram](image-url)

  - NEO locations:
    - Vancouver
    - Edmonton
    - San Jose
    - San Francisco
    - Minneapolis
    - Chicago
    - Boston
    - Summit
    - Albany
    - New York

  - NW locations:
    - New York
    - Albany

  - White budget:
    - $= 2$
    - $\geq 7$
GMDL Problem Statement
(spatial case)

• $k$ totally ordered dimensions $D_i \rightarrow S$ (set of all cells)
• $B$ (blue) and $R$ (red) – colored cells
• $W = S - (B \cup R)$ (white cells)
• Find axis-parallel hyper-rectangles $\{R_1, \ldots, R_m\}$ (i.e., GMDL covering) s.t.:
  – $(R_1 \cup \ldots \cup R_m) \cap R = \emptyset$ (validity)
  – $|(R_1 \cup \ldots \cup R_m) \cap W| \leq w$ (white budget)
  – $m$ is the least possible (optimality)
(G)MDL Problem Statement
(hierarchical case)

- k (tree) hierarchical dimensions
- cell = tuple of leaves
- region = tuple of nodes
- region R covers cell c iff c is a descendant of R, component-wise
- covering rules similar to spatial case
- MDL/GMDL problem formulations analogous
Algorithms for spatial GMDL

- challenges for spatial: even MDL 2D is NP-hard, so we must turn to heuristics
- important properties:
  - blue-maximality
  - non-redundancy
- Algorithms for spatial GMDL:
  - bottom-up pairwise (BP) merging
  - R-tree splitting (RTS) [based on Garcia+98]
  - color-aware splitting (CAS)
  - CAS corner
Algorithms for spatial GMDL (CAS)

• build indices $I_R$, $I_B$ for red and blue cells
• start with $C = \text{region } R \text{ covering all blue cells;}$
  \text{curr-consum } = \# \text{ white cells in } R$
• while ($\exists R \in C \text{ containing a red cell} \}$ {
  – grow the red cell to a larger blue-free region (using $I_B$)
  – split R into at most 2k regions (excluding the grown red region)
  – replace R by new regions }$
• while (curr-consum > w) \}$ {
  – split as above, but based on white cells }$
• return $C$
CAS – An Example

- trade-off
  - non-overlapping regions ➞ loss in quality
  - overlapping regions ➞ greater bookkeeping overhead

- Algorithms RTS, the two CAS’ ➞ non-redundant valid/feasible solutions
- BP ➞ may produce redundant solution; can be made non-redundant
Categorical Case – MDL

- ∃ key diff. between spatial and categorical?
- optimal covering → non-redundant
- optimal need *not* be blue-maximal, but can be expanded into one
- is blue-maximal non-redundant MDL covering unique? what about their size?
A spatial example

two blue-maximal non-redundant coverings of diff. size
Categorical – fundamentals

• projection of regions on dimensions: e.g., (MW, women’s) – projection on location = {chicago, minneapolis}.

• **Claim:** R, S any categorical regions (tree hierarchies); $R_i$ – projection of R on dimension i; $\forall i$, $R_i \subseteq S_i$ or $S_i \subseteq R_i$ or $R_i \cap S_i = \emptyset$

• see violation in “tough” spatial example

• major factor in deciding complexity
Theorem: space of $k$ categorical dimensions with tree hierarchies $\rightarrow$ unique blue-maximal non-redundant MDL covering.

Corollary: (i) the said covering can be obtained on a per hierarchy basis. (ii) furthermore, it can be done in polynomial time.
**Categorical case – MDL algorithm illustrated**

Propagate

Initialize

Before redundancy check

After redundancy check
Categorical case – MDL

- Lemma: Optimal MDL covering for a categorical space with tree hierarchies can be obtained by visiting each node once and each node of last hierarchy twice.
- Key idea: for tree hierarchies, finding all blue-maximal regions and removing redundant ones yields the optimal covering.
**Categorical case – GMDL**  

- Basic idea: for each internal node, determine the cost and gain of involving it in a GMDL covering; sort candidates in decreasing gain order and increasing cost. Pick greedily.

- Example:

<table>
<thead>
<tr>
<th>candidate</th>
<th>(1,h)</th>
<th>(2,h)</th>
<th>(3,h)</th>
<th>(4,h)</th>
<th>(5,h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>occurrence</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>max-gain</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>cost</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>X</td>
<td>3</td>
</tr>
</tbody>
</table>
Categorical Case – GMDL (contd.)

- Compile similar info. for other parents of leaves; sort and pick best w cells for color change. [drop candidates with cost X or 0.]
- Run MDL on the new data.
Related Work

- Substantial work on using MDL for summarization principle in data compression [Ristad & Thomas 95], decision trees [Quinaín & Rivest 89, Mehta+ 95], learning of patterns [Kilpelinien 95], etc.
- Summarizing cube query answers and (G)MDL on categorical spaces – novel.
Summary & Future Work

- summarization using MDL/GMDL as a principle
- MDL on spatial – NP-complete even on 2D; utility of GMDL – trade compactness for quality (i.e., include “impurity” in answers)
- Heuristic algorithms
- Efficient algo. for MDL for categorical with tree hierarchies
- Heuristics for GMDL
- Experimental validation
Future Work

• What is the best we can do to summarize data with both spatial and categorical dimensions?
• How far can we push the poly time complexity? (e.g., almost-tree hierarchies? Can we impose restrictions on “allowable” intervals even on spatial dimensions?)