Generalized Expansion Dimension

[PTDM 2012]

Michael E. Houle[1] Hisashi Kashima[2] Michael Nett[1,2]

[1] National Institute of Informatics, Japan

[2] The University of Tokyo, Japan

東京大学
THE UNIVERSITY OF TOKYO

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Outline

Motivation

Dimensionality tells us how **difficult** our data is to work with (curse of dimensionality).

- Often misunderstood in terms of **representational dimensionality**.
- More appropriately addressed by **intrinsic dimensionality**.

Why do we care in **practice**?

- Analysis of cost of **fundamental operations** in data mining.
- Design of **efficient heuristics**.
- Support **algorithmic decisions** made at runtime.

Modeling Intrinsic Dimensionality

Counting **numbers of features** is a simple way to model dimensionality, but it is often **not appropriate**.

Different models of **intrinsic dimensionality** have been proposed:

- Fractal Dimension
- Aspect Ratio
- Covering Dimension
- Disorder Inequalities
- Expansion Dimension (Karger & Ruhl)

Generalized Expansion Dimension (Core Idea)

Look for the dimensionality that best explains the observed density of data around a reference point.

Modeling Intrinsic Dimensionality

Key Observation:

- Neighborhood volumes are determined by radii and dimensionality.

Example: Euclidean space $(\mathbb{R}^m, d)$.

$$\lambda(B(q,r)) = r^m \cdot \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)}$$

Modeling Intrinsic Dimensionality

General approach:

1. Measure **neighborhood volumes** and **radii**.
2. Solve for **dimension**.

Examples:

- Euclidean: Testing with two neighborhoods yields

$$\frac{\lambda(B(q,r_1))}{\lambda(B(q,r_2))} = \left(\frac{r_1}{r_2}\right)^m \Leftrightarrow m = \frac{\log \lambda(B(q,r_1)) - \log \lambda(B(q,r_2))}{\log r_1 - \log r_2}$$

- Different results for **Hamming** and **vector angle** distance.

Modeling Intrinsic Dimensionality

Let $S \subseteq \mathbb{U}$ be a point set. Neighborhood measures are **estimated by numbers of points of S captured**.

$$\lambda(B(q,r)) \sim |B(q,r) \cap S| \Rightarrow \boxed{m \sim \frac{\log k_1 - \log k_2}{\log r_1 - \log r_2}}$$

locality aware

uneven growth

Uneven growth: Our volume estimates do **not** grow smoothly ($r \in \mathbb{R}$, but $k \in \mathbb{N}$).

- Different choices of r provide different estimations of m.
- We use **medians of medians** (high stability, different percentiles are possible, ...).

Modeling Intrinsic Dimensionality

Algorithm. Dimensionality testing at query location $x \in \mathbb{U}$ with neighborhood sizes in the range $k_-, \ldots, k_+$.

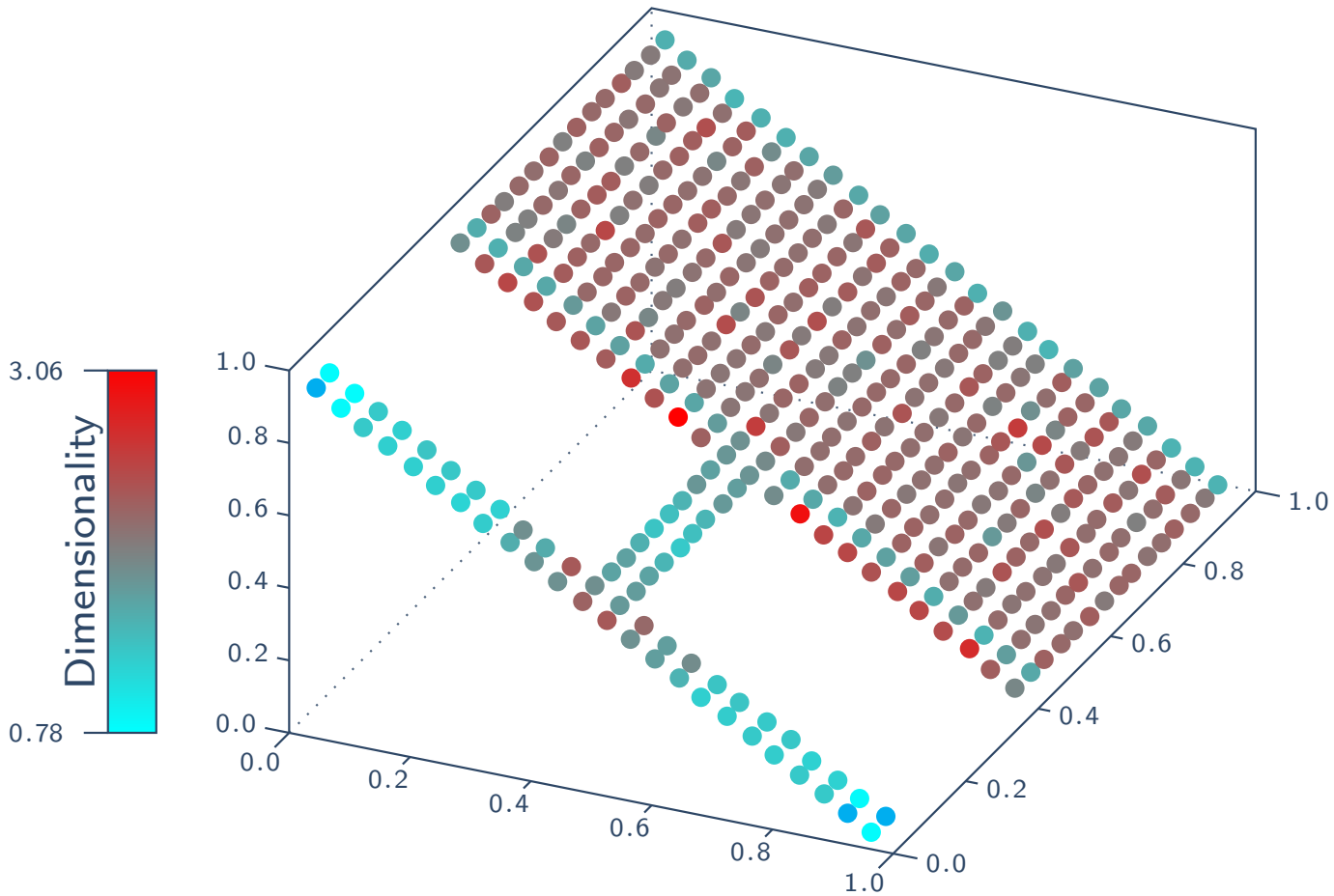
1. Let $Q = (q_{k_-}, \ldots, q_{k_+})$ be the sorted list of query-distances (of neighbors ranked between k_- and k_+).
2. Let $K = \{k_-, \ldots, k_+\}$ be the range of considered neighborhood sizes. For any choice of $k \in K$, let $A_k \triangleq \{(k, i) \ : \ i \in K \text{ and } i \neq k\}$.
3. For any $k \in K$ let m_k be the median of the individual tests involving one neighborhood of size k, that is,

$$m_k = \underset{(k,i) \in A_k}{median}\, \Delta(B(x, q_k), B(x, q_i) | \, S, \mathbb{U}, D).$$

4. Report $median_{k \in K}\, m_k$ as the dimensionality at x.

Modeling Intrinsic Dimensionality

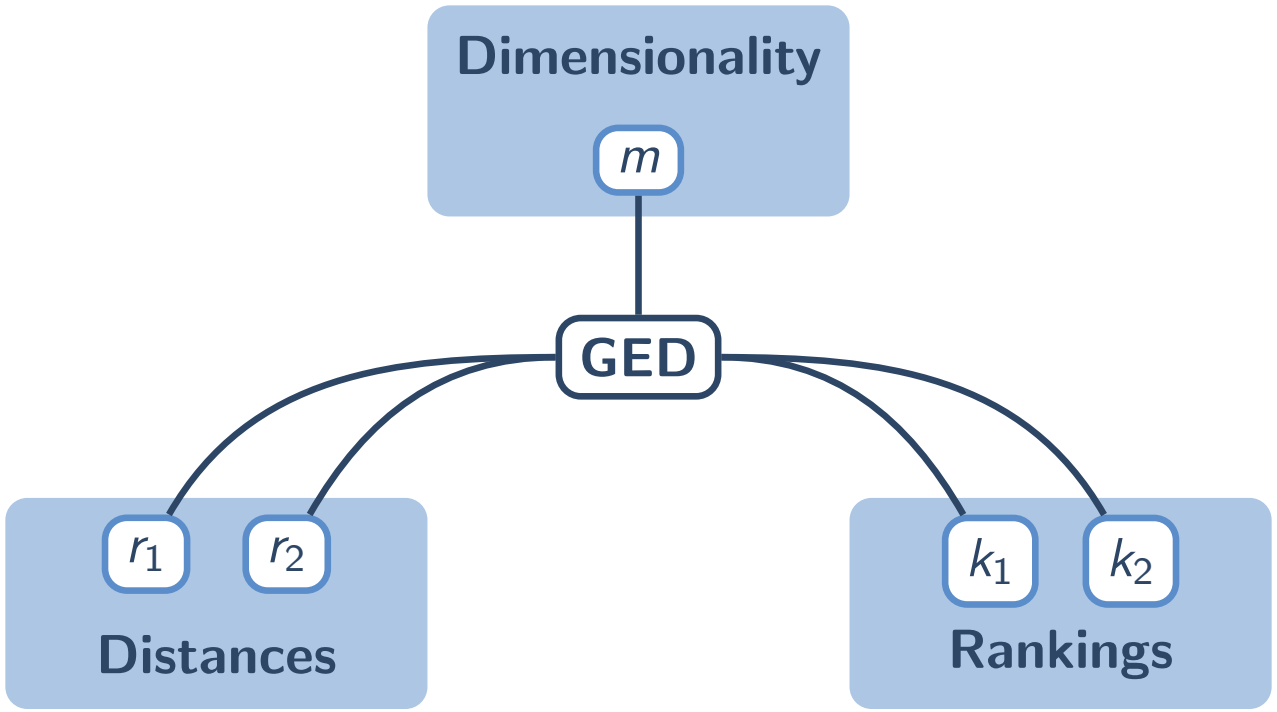
- Visualization of the local intrinsic dimensionality on a perforated plane in $\mathbb{R}^3$:



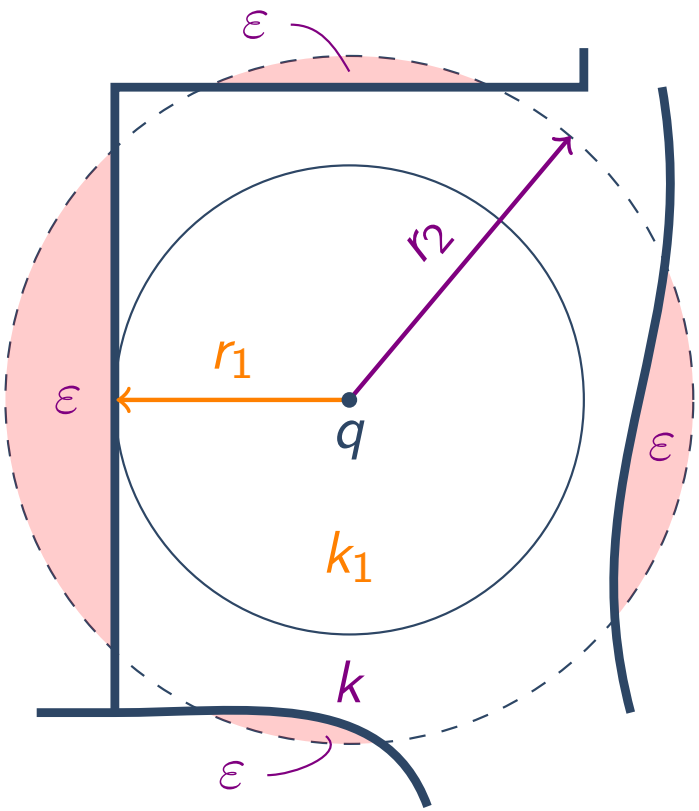
Applications

The Big Picture:

- If sufficiently many variables are known, we can **bound unknown variables**.



Dimensional Pruning for Similarity Search



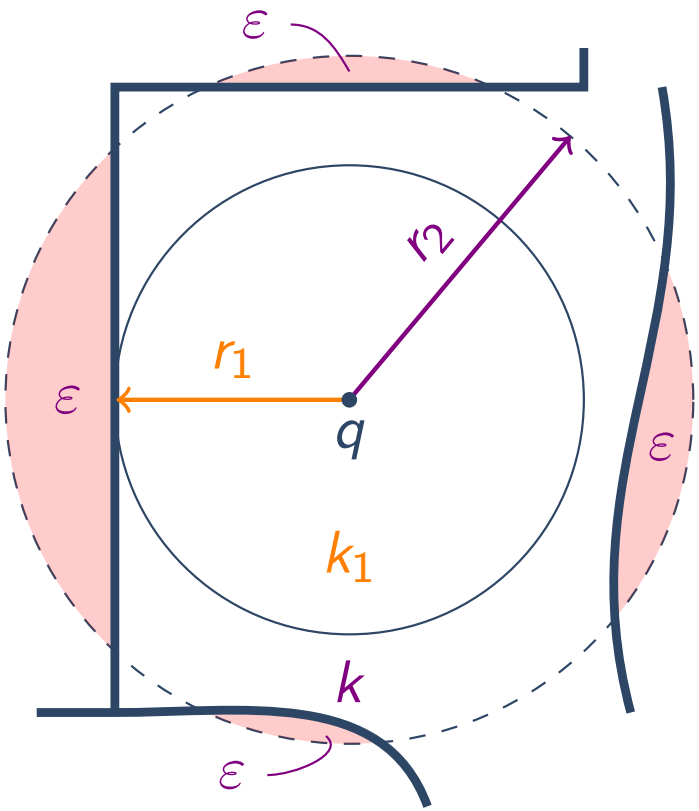
Dimensional test with neighborhoods centered at q:

- Outer ball containing k candidates ($d_k(q) \leq r_2$).
- Inner ball tangent to most restrictive (closest) separator.

Observations:

- Inner ball is part of the solution.
- If $\varepsilon < 0$ we can prune.

Dimensional Pruning for Similarity Search



- Maximum GED ($\Delta_{\max}$) at least

$$\frac{\log(k+\varepsilon) - \log k_1}{\log r_2 - \log r_1}$$

- Can safely terminate if

$$k_1 \left(\frac{r_2}{r_1}\right)^{\Delta_{\max}} < k+1$$

- Heuristic: Estimate $\Delta_{\max}$ by sampling and pick a percentile.

Applications

In practice, direct computation of m can be **too expensive**.

Solution: Estimate order-statistic quantities (percentiles, maxima) of dimensionality values by **sampling reference points**.

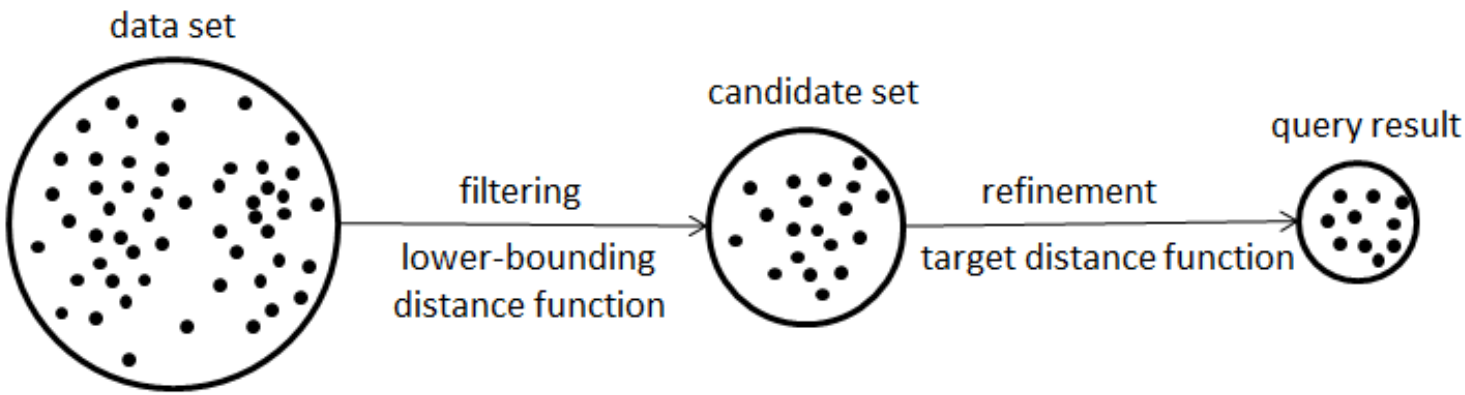
- Percentiles often correspond to algorithmic performance (recall rates, . . .).
- Estimates concentrate sharply around their true values.

Multi-Step Similarity Search

Clustering II, 10:40 – 11:00, Room: Salle des Nations I — "Dimensional Testing for Multi-Step Similarity Search", M. E. Houle, X. Ma, M. Nett, and V. Oria

The situation:

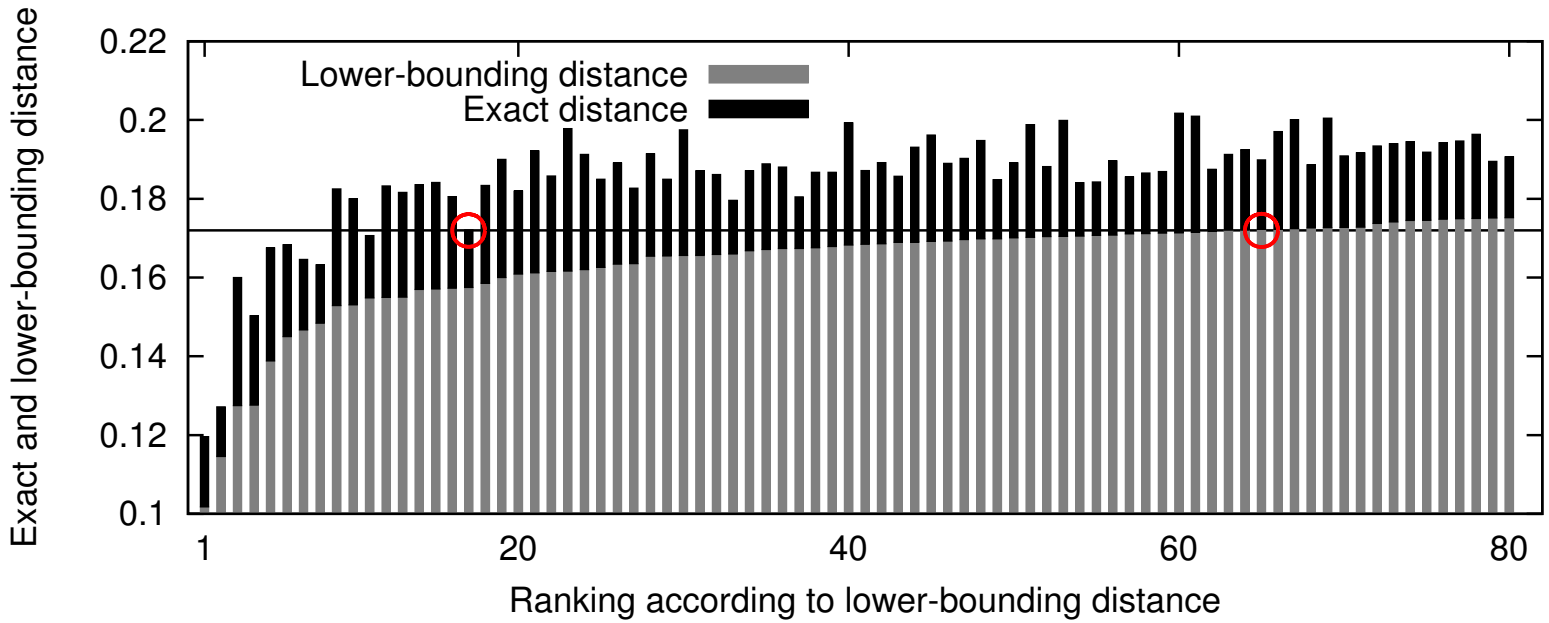
- We have a **lower-bounding distance** $d'(x, y) \leq d(x, y)$. (filtering, feature selection, ...)



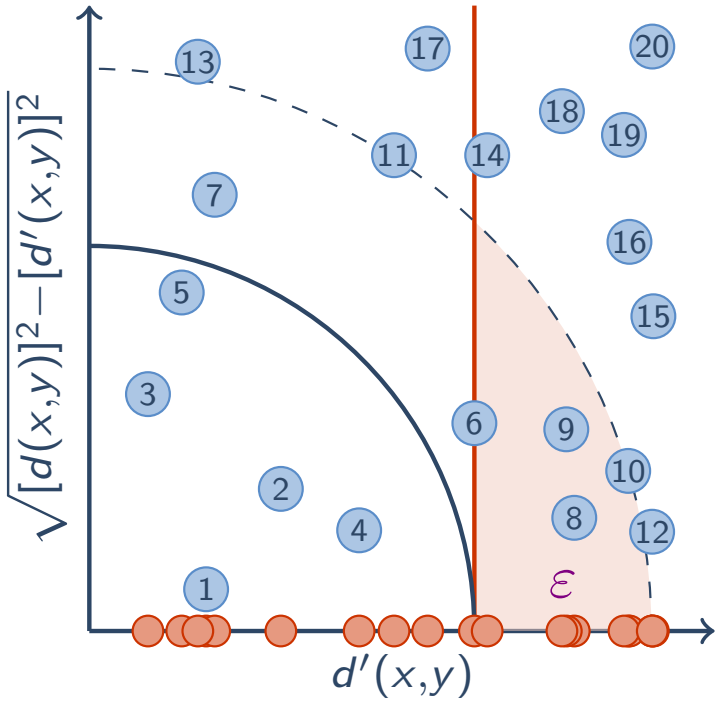
- Special case of **ranked list aggregation**.

Multi-Step Similarity Search

- **Optimal solution** assuming no knowledge on future distance values (Seidl & Kriegel, 1998).
- Potential for early termination:



Multi-Step Similarity Search



- Put conditions on ε being zero.
- Solution parameterized in t.
- Theoretically guaranteed correctness if $t \geq \Delta_{\max}$.
- Competitive heuristic for smaller choices of t.

Publications

- D. R. Karger and M. Ruhl. "*Finding Nearest Neighbors in Growth-Restricted Metrics*", STOC 2002.
- A. Beygelzimer, S. Kakade, and J. Langford, "*Cover Trees for Nearest Neighbor*", ICML 2006.
- M. E. Houle and M. Nett, "*Rank-Based Similarity Search: Reducing the Dimensional Dependence*", Tech. Rep. NII-2012-004E.
- M. E. Houle, S. Chawla, and T. deVries, "*Finding Local Anomalies in Very High Dimensional Space*", ICDM 2010.
- M. E. Houle, X. Ma, M. Nett, and V. Oria, "*Dimensional Testing for Multi-Step Similarity Search*", ICDM 2012.

Challenges

Concerning the **model**.

- Other domains (set lattices, scale-free networks, . . .).
- General model of complexity of data.
- Comparison of different models (GED vs. PCA, network scale, . . .).

Current and future work:

- Dimensional pruning
- Anomaly detection
- Feature selection
- Cost balancing / accuracy balancing