Generalized Expansion Dimension

Michael E. HOULE [1] **Hisashi KASHIMA** [2] **Michael NETT** [1,2]

[1] National Institute of Informatics, Japan [2] The University of Tokyo, Japan

What?

- Measure-based framework for estimation of intrinsic dimensionality of point sets embedded into a feature space.
- Particular interpretations of the model with respect to popular domains and similarity functions (L_p-norms, cosine similarity, and Hamming distances).
- Automated assessment of local and global complexity of data.
- Application of dimensionality values to guide algorithmic decisions at runtime.

Why?

- Complexity of data is typically assessed in terms of its *representational* dimensionality.
- The '*curse of dimensionality*' hinders many data mining applications as the complexity of data grows.
- Large body of empirical and theoretical evidence for this phenomenon.
- However, the effect on applications is often not as severe as predicted by theory.
- Representational dimensionality may not be an appropriate measure of complexity.

Notation

Let $(\mathbb{U}, D, \lambda)$ be a space with distance metric

$$D : \mathbb{U} \times \mathbb{U} \to \mathbb{R}_{\geqslant 0}$$

and measure

$$\lambda : \mathcal{P}(\mathbb{U}) \to \mathbb{R}_{\geqslant 0}.$$

Our model is based on observing measures $\lambda(B(x,r))$ of closed neighborhoods of radius r, centered at $x \in \mathbb{U}$, where

$$B(x,r) \triangleq \{ y \in \mathbb{U} \ : \ D(x,y) \leqslant r \}.$$

Thought Experiment

- Suppose that we know the nature of the underlying feature space $\mathbb{U}$. One possible example: the Euclidean space $\mathbb{R}^m$.
- However, the dimension m is hidden from us.
- Can we determine the value of m, given access to an oracle computing the measure of any neighborhood $B(x,r)$?

Since we know that the measure of $B(x,r)$ in m-dimensional Euclidean space is

$$\lambda(B(x,r)) = \int_{B(x,r)} \mathrm{d}x = \frac{\sqrt{\pi^m}}{\Gamma(\frac{m}{2}+1)} \cdot r^m,$$

we consider the ratio of measure of two neighborhoods with distinct radii $r_1 \neq r_2$:

$$\frac{\lambda(B(x_1,r_1))}{\lambda(B(x_2,r_2))} = \cdots = \left(\frac{r_1}{r_2}\right)^m.$$

Solving for m provides

$$m = \frac{\log \lambda(B(x_1,r_1)) - \log \lambda(B(x_2,r_2))}{\log r_1 - \log r_2}.$$

Finite Point Sets

- Suppose that we populate a given region X of our space by a point set $S \subseteq X$ of density ρ, selected uniformly at random.
- We may think of an individual sample x as a condensed representative of its Voronoi region with respect to S.
- Taking an arbitrary neighborhood $B(x,r)$, we would expect its measure with respect to both X and S to be roughly proportional.
- Moreover, we expect both measures to converge as the resolution of the sample S becomes finer.

A similar line of reasoning applies when we let a finite data set $S \subseteq \mathbb{U}$ serve as a representative of the space $\mathbb{U}$ from which it is drawn.
Here, the measure of a region X would be assessed with respect to S rather than $\mathbb{U}$. Since S is finite, a combinatorial measure is needed:

$$\lambda_S(X) = |S \cap X|.$$

For the case when X is a neighborhood, measure queries can be answered using a suitable similarity search structure for point set S.

Interpretation for L_p-norms

The Minkowski L_p-norm on $\mathbb{R}^m$ is defined as

$$\|x\|_p \triangleq \left(\sum_{i=1}^{m} x_i^p \right)^{\frac{1}{p}}.$$

Common examples include the Euclidean norm ($p = 2$), the Manhattan norm ($p = 1$) and the maximum norm ($p \to \infty$). For $p \geqslant 1$, the function $D_p(x,y) = \|x - y\|_p$ is a distance metric.
The measure of $B(x,r)$ in $(\mathbb{R}^m, D_p, \lambda)$ is given by

$$\lambda(B(x,r)) = \int_{B(x,r)} \mathrm{d}x = \left(\frac{2r}{p}\right)^m \int_{\|x\|_2 \leqslant 1} \prod_{i=1}^{m} x_i^{\frac{2}{p}-1} \mathrm{d}x$$
$$= \frac{r^m}{p^m} \cdot \frac{2^m \cdot \Gamma(\frac{1}{p})^m}{\Gamma(1+\frac{m}{p})}.$$

We can estimate the intrinsic dimensionality of a point set by relating the measure of two neighborhood sets of different radii

$$\Delta(B(x_i,r_i), B(x_j,r_j) \ : \ S, \mathbb{R}^m, D_p)$$
$$= \frac{\log |S \cap B(x_i,r_i)| - \log |S \cap B(x_j,r_j)|}{\log r_i - \log r_j}.$$

Bit Strings with Hamming Distance

Let $\mathbb{U} = \{0,1\}^m$ be the space of all bit strings of length m. The *Hamming distance* (or *mismatch distance*) on $\mathbb{U}$ is defined as

$$D_H(x,y) \triangleq |\{ 1 \leqslant i \leqslant m \ : \ x_i \neq y_i \}|.$$

We can *approximate* the intrinsic dimensionality of a point set $S \subseteq \mathbb{U}$ by

$$\Delta(B(x,r), B(x,r-1) \ : \ S, \{0,1\}^m, D_H)$$
$$\approx \ r! \cdot (\lambda(B(x,r)) - \lambda(B(x,r-1)))^{\frac{1}{r}} + \frac{r-1}{2}$$

with a relative error no more than

$$\frac{r-1}{2\Delta(B(x,r), B(x,r-1) \ : \ S, \{0,1\}^m, D_H}.$$

Interpretation for Cosine Similarities

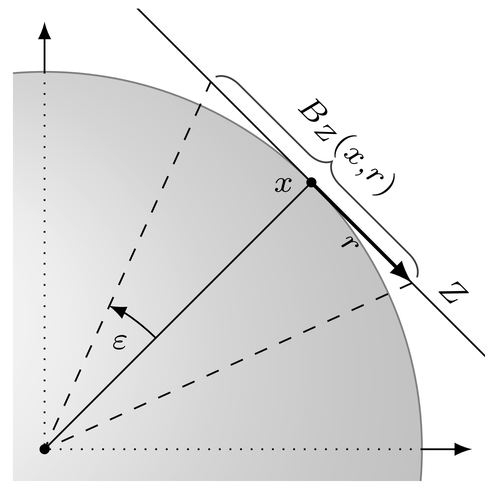
In practice, *vector angle metric* (or equivalently *cosine similarity*) is used to compute similarities between pairs of documents expressed as keyword vectors. The metric is defined on $\mathbb{R}^m_+ \subseteq \mathbb{R}^m$ as follows:

$$D_\angle(x,y) \triangleq \cos^{-1}\left(\frac{x^\top y}{\|x\|_2 \cdot \|y\|_2}\right).$$

The dimensionality can be computed as

$$\Delta(B(x_i,r_i), B(x_j,r_j) \ : \ S, \mathbb{R}^m_+, D_\angle)$$
$$= \frac{\log |S \cap B(x_i,r_i)| - \log |S \cap B(x_j,r_j)|}{\log \tan r_i - \log \tan r_j},$$

subject to $r_i \neq r_j$.



Algorithm For Dimensionality Testing

Let $x \in \mathbb{U}$ be the neighborhood center with respect to which we want to estimate the intrinsic dimensionality.

- Let $K = \{k_-, \ldots, k_+\}$ be the range of considered neighborhood sizes.
- Let $Q = \{r_{k_-}, \ldots, r_{k_+}\}$ contain the query-distances of neighbors ranked in the range K, with respect to x.
- Following the convention that division by zero produces a value of $+\infty$, return the median of medians across all considered neighborhood sizes:

$$\operatorname*{med}_{i \in K} \left\{ \operatorname*{med}_{j \in K} \left\{ \Delta(B(x,r_i), B(x,r_j) \ : \ S, \mathbb{U}, D) \right\} \right\}$$

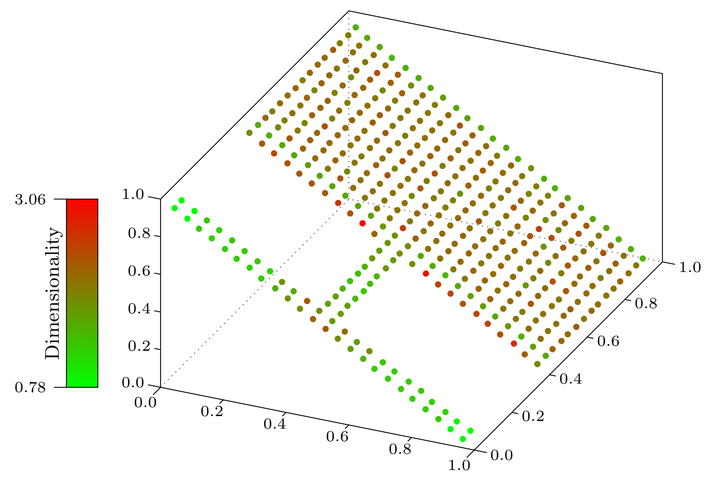


Figure: Local dimensionality values in a point set in $(\mathbb{R}^3, D_2)$ arranged near the surface of a plane, over the neighborhood range $K = \{10, \ldots, 20\}$.

Additional Material

- typhoon.nii.ac.jp/ged
- Preprint
- Poster
- Implementation (C++)

連絡先: Michael E. HOULE (フール マイケル) / 国立情報学研究所 客員教授
☎: 03-4212-2538 FAX: 03-3556-1916 ✉: meh@nii.ac.jp