Dimension Induced Clustering

Aris Gionis
Alexander Hinneburg
Spiros Papadimitriou
Panayiotis Tsaparas

HIIT, University of Helsinki
Martin Luther University, Halle
Carnegie Melon University
HIIT, University of Helsinki

Introduction

Vector Space Motivation

- Observation: data points cannot fill the space =>
  Data lie on one or more low-dimensional manifolds
- Real data exhibit patterns and regularities
Example

- The red points form a 1d manifold in the 2d space.

- A low dimensional manifold must contain sufficient number of points that are densely packed
  - density-based methods?

Dimension Induced Clustering

- How to separate river and lake
  - River and lake have same density
  - Both are spatially connected
  - But they differ in dimensionality

- Density is still necessary for separating lake from surroundings
Dimension Induced Clustering

- Problem
  - Given a set of data objects with a distance function
  - Find dense subsets of objects with similar dimensionality

Other Applications

- Indexing
  - efficient approximation of nearest neighbor for metric data,
    assumes bounded intrinsic dimensionality
    [Krauthgamer & Lee, ICALP 2004]

- Mixture Models of PCA
  - needs average dimensionality as parameter
    [Aggarwal & Yu, SIGMOD 2002], [P. Agarwal et al, PODS 2004]
What is dimension?

- Approach using **representation**
  - dimension is the number of coordinates
  - decompose data space into set of linear sub-spaces densely filled with points
- Drawbacks
  - assumes vector spaces
  - only linear manifolds

What is dimension?

- Approach using relative distances
  - use distances between objects only
  - extend notion of **fractal dimension**
- Advantages
  - complex curved manifolds
  - applicable to metric spaces
Fractal Dimension

- Correlation Dimension
  - Set of objects $X$, $|X|=n$
  - Distance function $d : X \times X \rightarrow \mathbb{R}$
  - Points in the ball of radius $r$ around $x$
    $B(x, r) = \{ y \mid y \in X, d(x, y) \leq r \}$
  - Correlation Integral
    $C(r) = \lim_{n \to \infty} \frac{1}{n} \sum_{x \in X} \frac{|B(x, r)|}{n}$
  - Correlation Dimension
    $d_{corr} = \lim_{r, r' \to 0} \frac{\log C(r) - \log C(r')}{\log r - \log r'}$

Intuition behind definition

- $d_{corr} = 1$
- $O(r)$
- $|B(x, r)|$
- $r$
- $d_{corr} = 2$
- $O(r^2)$
- $|B(x, r)|$
- $r$
Fractal Dimension

- In real life, datasets are finite
  \[
  C(r) = \frac{1}{n} \sum_{x \in X} \frac{|B(x,r)|}{n}
  \]

- Calculation of correlation dimension:
  fit a line on the log-log plot of \( C(r) \)
**Fractal Dimension**

- What if the data is non-homogeneous?

**Local Fractal Dimension**

- However, looking at A and B individually
Local Fractal Dimension

- Local Growth Curve
  \[ G_x(r) = \lim_{n \to \infty} \frac{1}{n} |B(x, r)| \]
- Local Correlation Dimension
  \[ d_x = \lim_{r, r' \to 0} \frac{\log G_x(r) - \log G_x(r')}{\log r - \log r'} \]
- For finite data \[ G_x(r) = \frac{1}{n} |B(x, r)| \]
  \( d_x \) is estimated by fitting a line

Local Fractal Dimension

- Linear Growth Model of an object \( x_i \)
  \[ L_{x_i}(\log r) = d_i \log r + b_i \]
Linear Growth Model

\[ L_{x_i}(\log r) = d_i \log r + b_i \]

- \( d_i \): rate of growth of \( \log G_x(r) \) – dimensionality
- \( b_i \): value of \( L_x(\log r) \) at radius 1 – density
- \( L_x(\log r^*) \): density at radius \( r^* \)

The model can be summarized with two values: \( d_i \) and \( c_i = L_x(\log r^*) \)
  - how do we select \( r^* \) ?

Selecting \( r^* \)

- Idea: choose \( r^* \) such that \( c_i \)'s and \( d_i \)'s are un-correlated

Lemma 1. The choice of \( r^* \) for which \( d_i \) and \( c_i \) are un-correlated is given by

\[
\log r^* = -\frac{\sum (d_i - \bar{d})(b_i - \bar{b})}{\sum (d_i - \bar{d})^2}
\]
Local Representation

- Local Representation of point $x_i$
  
  \[
  l(x_i) = (d_i, c_i)
  \]
  \[
  c_i = L_{x_i}(\log r^*)
  \]

- Captures the view of the world for each point

The fitting interval

\[
L_{x_i}(\log r) = d_i \log r + b_i
\]

- The linear growth model is defined over a subset of the neighbors of $x$
  - Clipping from above
  - Clipping from below
Algorithm

Algorithm 1 The DIC algorithm

Input: Dataset $X$ of $n$ points, number of clusters $b$
Output: Clustering of $X$ into $b$ clusters

1: for all $i \in \{1, \ldots, n\}$ do
2: Compute $k$-th NN of $x_i$, for $k = k_{\min} \ldots k_{\max}$
3: Compute the local representation $(d_i, c_i)$ of $x_i$.
4: end for
5: $X_{LR} = \{(d_1, c_1), \ldots, (d_n, c_n)\}$
6: Cluster the set $X_{LR}$ into $b$ clusters.

Experiments

- Detection of m-flats in high-dimensional space

(a) 2d flat in 3d space
Classification error = 8.1%

(a) 40d flat in 50d space
Classification error = 1.2%
Comparison to Optics

- Optics: density based hierarchical clustering

Low Rank Sub-Matrix

- Combinatorial low-rank sub-matrix in a random Matrix
- Apply DIC to set of columns and set of rows
- Final Clusters are the Cartesian product of row and column clusters
Experiments

- Gene Expression Data (gene clustering)

Yeast data from George Church, Harvard

Neither density nor dimensionality alone can detect the structure
Conclusion

- Find subsets with low fractal dimensionality
- Local Representation
  - local fractal dimensionality
  - local density
- Visualization of the cluster structure