Interdependence of clusters measures and distance distribution in compact metric spaces

Pushnyakov Alexey

MIPT

IDP-11, Barcelona, 2016

Motivation

What is a *good* metric like?

- compactness principle: close objects should be in the same class rather than in different ones
- there are specific intra-cluster and inter-cluster distances
- metric space is a disjoint union of clusters separated one from each other
- if there are exactly k clusters then among every $k + 1$ points there exists two points form the same class

Problem statement

A compact metric space (X, ρ) with a Borel measure μ is considered.

Defenition

A measurable set of diameter at most r is called r-cluster.

Defenition

A family of $2r$-clusters $\mathcal{X} = \{X_1, \dots X_k\}$ is called r-cluster structure of order k if $\rho(X_i, X_j) \geqslant r$ for all $1 \leqslant i < j \leqslant k$, where $\rho(A, B) = \inf\{\rho(x, y) \colon x \in A,\ y \in B\}$. By measure of cluster structure $\mathcal{X}$ we mean value $\mu(\mathcal{X}) \stackrel{def}{=} \sum_{i=1}^{k} \mu(X_i)$.

Statement

There exists a r-cluster structure of order k of maximum measure.

Let $\mathcal{X}^*$ be a r-cluster structure of order k of maximum measure.
If metric is *good* we have $\mu(\mathcal{X}^*) \approx \mu(X)$.
What restriction should we impose to guarantee that $\mu(\mathcal{X})$ *is close to* $\mu(X)$*?*

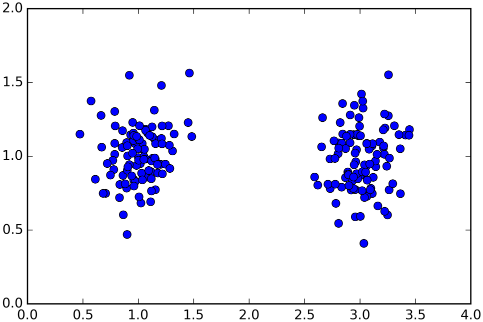
$$\rho(x, y) \in \begin{cases} [0, r], & \text{short edge;} \\ (r, 3r], & \text{medium edge;} \\ (3r, +\infty), & \text{long edge.} \end{cases}$$

Anticlique of order k is a set k points such that there are not short edges between them.

$$\mu\{(x, y) \in X^2 \colon r < \rho(x, y) \leqslant 3r\} \leqslant \alpha\mu(X)^2$$

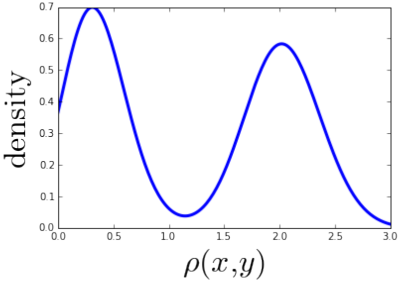
$$\mu\{(x_1, \ldots, x_{k+1}) \in X^{k+1} \colon \rho(x_i, x_j) > r,\ 1 \leqslant i < j \leqslant k+1\} \leqslant \beta\mu(X)^{k+1}$$

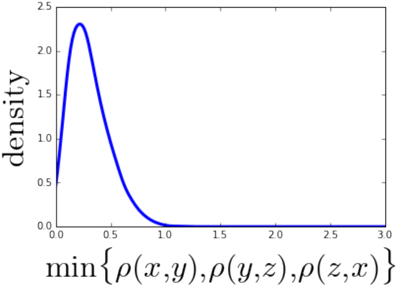
Model example



$X = X_1 \sqcup X_2$
$X_1 \sim \mathcal{N}(m_1, \sigma^2 I)$
$X_2 \sim \mathcal{N}(m_2, \sigma^2 I)$
$\sigma = 0.2,\ m_1 = (1, 1)^T,\ m_2 = (3, 1)^T$

Model example





Distributions of $\rho(x, y)$ and $\min\{\rho(x, y), \rho(y, z), \rho(z, y)\}$ have features described above.

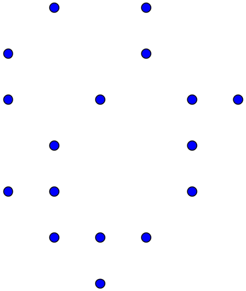
Greedy cluster structure

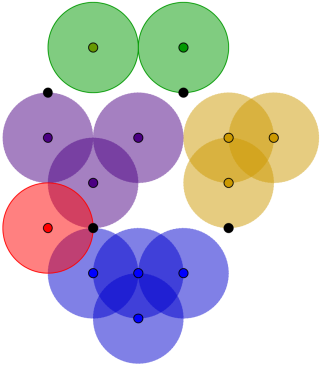
Let a metric space X be finite and μ is uniform measure.

- $Z_1, \ldots, Z_m$ are pairwise disjoint sets.
- Let X_{m+1} be $2r$-cluster of maximum cardinality in $X \setminus \bigcup\limits_{i=1}^{m} Z_i$.
- Z_{m+1} is r-neighborhood of X_{m+1} in $X \setminus \bigcup\limits_{i=1}^{m} Z_i$:

$$Z_{m+1} = \left\{ x \in X \setminus \bigcup_{i=1}^{m} Z_i \colon \rho(x, X_{m+1}) < r \right\}$$

Greedy cluster structure





Defenition

The partition $X = \bigsqcup_{i=1}^{n} Z_i$ is called a greedy cluster partition and the family of $2r$-clusters $\{X_1, \ldots, X_k\}$ is called a greedy r-cluster structure of order k.

Goal is to get the following bound

$$\sum_{i=1}^{k} |X_i| = (1 + o(1))|X|, \alpha + \beta \to 0$$

Main steps

Let σ be a permutation such that $|Z_{\sigma(1)}| \geqslant |Z_{\sigma(2)}| \geqslant \ldots$ and by defention $W_i = |Z_{\sigma(i)}|$.

- $\sum_{i=1}^{k} W_i \geqslant (1 + o(1))|X|$
- $\sum_{i=1}^{k} (W_i - |X_{\sigma(i)}|) = o(1)|X|$
- generalize bound for a compact metric space.

Let $T_s(i_1, \ldots, i_s)$ be number of r-anticliques of order s intersecting sets Z_{i_j} by exactly one point.

Statement

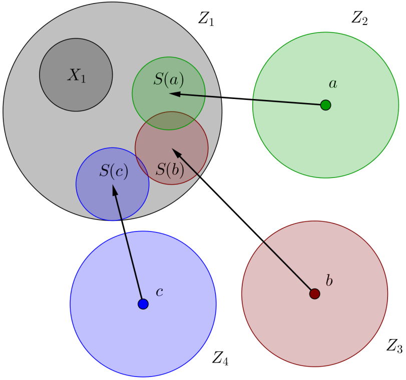
Assume $i_1 < \ldots < i_s$ and $s \geqslant 2$ then

$$T_s(i_1, \ldots, i_s) \geqslant \frac{|Z_{i_1}|}{s} T_{s-1}(i_2, \ldots, i_s)$$

$$T_s(i_1, \ldots, i_s) \geqslant \frac{1}{s!} \prod_{j=1}^{s} |Z_{i_j}|$$

$$\sigma_s(y_1, \ldots y_n) \stackrel{\text{def}}{=} \sum_{1 \leqslant i_1 < \ldots < i_s \leqslant n} \prod_{j=1}^{s} y_j,$$

$$\sigma_{k+1}(W_1, \ldots, W_n) \leqslant \beta |X|^{k+1}$$



$$S(a) = \{x \in Z_1 \colon \rho(x, a) \leqslant r\}$$

Lower bound for $\sum_{i=1}^{k} W_i$

We obtain following optimization problem:

$$\begin{cases} f(\mathbf{w}) = \sum\limits_{j=1}^{k} w_j \to \min\limits_{\mathbf{w}} \\ w_i \geq 0 \\ w_i \geq w_j, \ i \leq j \\ \sum\limits_{i=1}^{n} w_i = 1 \\ \sigma_{k+1}(w_1, \dots, w_n) \leqslant c \end{cases} \tag{1}$$

Statement

If $\mathbf{w}$ *is a solution of (1) and* $w_k = \lambda > 0$ *then* $\mathbf{w} = (w_1, \underbrace{\lambda, \dots, \lambda}_{s}, \mu, 0, \dots, 0)$,

where $s \geqslant k-1$ *and* $\mu \ < \ \lambda$.

$$\sum_{i=1}^{k} W_i \geqslant |X| \left(1 - (k+1)\beta^{\frac{1}{k+1}}\right)$$

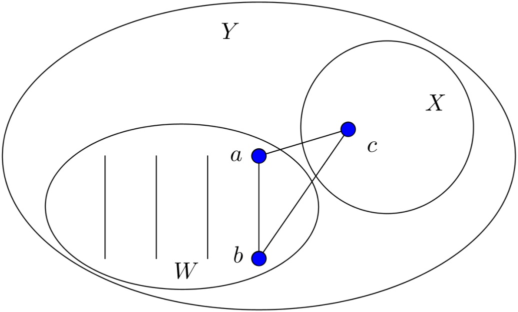
Inner structure of greedy cluster structure

Statement

Let (A, ρ) be a finite pseudometric space of diameter at most $3r$ and B is a $2r$-cluster of maximum cardinality then number of medium edges $M(A)$ no less than $\frac{1}{2}|A||A \setminus B|$.

- X_i is a $2r$-cluster of maximum cardinality in Z_i.
- Diameter of Z_i might be equal $4r$.
- Let M_i be a maximum matching of long edges covering the set $W_i \subset Z_i \setminus X_i$.
- By definition $Y_i = Z_i \setminus (X_i \cup W_i)$. Notice that $Y_i \cup X_i$ is a $3r$-cluster.

Inner structure of greedy cluster structure



$$\max\{\rho(a,c),\rho(b,c)\} > r$$

$$M(Z_i) \geqslant \frac{1}{2}(|X_i| + |Y_i|)|Y_i| + \frac{1}{2}|W_i||X_i|$$

Statement

Let $T_s(Z_i)$ be number of r-anticliques of order s in Z_i and $s \geqslant 3$ then

$$T_s(Z_i) \geqslant \frac{1}{s}(|Z_i| - (s-1)|X_i|)_+ T_{s-1}(Z_i)$$

- $I_1 = \{i \colon |X_i|(k+1) \leqslant |Z_i|\}$:

$$\sum_{i \in I_1} |Z_i| \leqslant ek\beta^{\frac{1}{k+1}}|X|$$

- $I_2 = \{i \notin I_1 \colon |Z_i| \geqslant \sqrt{\alpha}|X|\}$:

$$\sum_{i \in I_2} (|Z_i| - |X_i|) \leqslant \sqrt{\alpha}(k+1)|X|$$

Bound for measure of greedy cluster structure

Theorem

Let (X, ρ) *be a finite pseudometirc,* μ *is uniform measure on* X *and* $\mathcal{X}^*$ *is a r-cluster structure of maximum measure. If conditions*

$$\mu\{(x, y) \in X^2 \colon r < \rho(x, y) \leqslant 3r\} \leqslant \alpha\mu(X)^2$$

and

$$\mu\{(x_1, \ldots, x_{k+1}) \in X^{k+1} \colon \rho(x_i, x_j) > r\} \leqslant \beta\mu(X)^{k+1}$$

are satisfied then

$$\mu(\mathcal{X}^*) \geqslant \Psi(\alpha, \beta)|X|,$$

where

$$\Psi(\alpha, \beta) = 1 - \sqrt{\alpha}(2k+1) - (k(e+1)+1)\beta^{\frac{1}{k+1}}$$

Case of compact metric space

By given $\varepsilon > 0$ we construct a finite approximation $(X_\varepsilon, \rho_\varepsilon)$ for (X, ρ).

- Let $\bigsqcup_{i=1}^{N_\varepsilon} A_i$ be a partition of X into ε-clusters.
- $X_\varepsilon \stackrel{\text{def}}{=} \bigsqcup_{i=1}^{N_\varepsilon} B_i$

$$\rho_\varepsilon(x, y) = \begin{cases} 0, & x, y \in B_i \\ \rho(A_i, A_j), & x \in B_i, y \in B_j,\ i \neq j \end{cases}$$

- $|B_i|\mu(A_j) \approx |B_j|\mu(A_i)$

Theorem

Let (X, ρ) be a compact metric space, μ is Borel measure on X and $\mathcal{X}^*$ is a r-cluster structure of maximum measure. If conditions

$$\mu\{(x, y) \in X^2 \colon r < \rho(x, y) \leqslant 3r\} \leqslant \alpha\mu(X)^2$$

and

$$\mu\{(x_1, \ldots, x_{k+1}) \in X^{k+1} \colon \rho(x_i, x_j) > r\} \leqslant \beta\mu(X)^{k+1}$$

are satisfied then $\mu(\mathcal{X}^*) \geqslant \Psi(\alpha, \beta)|X|$