Outlier Detection

Sanaz Bahargam
Boston University

Evimaria Terzi
Boston University

A Big Perspective

We consider the problem of finding the outliers and partitioning the data to some groups, given the desired partition representative and we are allowed to remove a predefined number of outliers.

ℓ-Removal

Given $X = \{x_1, x_2, .., x_n\}$ and a vector τ, find ℓ point such that $\| mean(X \setminus X_\ell), \tau \|$ is minimized.

ℓ-Removal problem is NP-hard and NP-hard to approximate

Set Partitioning Problem < ℓ-Removal

Solving ℓ-Removal

Integer Regression: Find a 0-1 vector S such that $\|mean(X * S), \tau\|$ is minimized, and S contains at least n - ℓ 1s

$$\begin{bmatrix} 5 & 7 & 3 & 4 & 6 & 3 \\ 9 & 8 & 6 & 4 & 7 & 5 \\ 4 & 3 & 4 & 1 & 2 & 3 \\ 0 & 1 & 1 & 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix} = Xs \qquad S_i \in \{0,1\}$$

Step 1: Find a nonnegative real-valued vector S minimizing cost function

```
cvx_begin
    variable s(n) nonnegative
    minimize ( norm(data * s - mu , 2))
    subject to
        sum(s) >= 1
        for i=1:n
            s(i) <= 1 / (n - ℓ)
        end
cvx_end
```

Step 2: Transform S into an integer value vector which contains at least n - ℓ 1s

Target Partitioning

Given $X = \{x_1, x_2, .., x_n\}$ and vectors $\tau_1, \tau_2, \ldots, \tau_k$ partition X into k partitions such that $\Sigma \| \mu(C_i) - \tau_i \|_{(i=1..k)}$ is minimized.

Target Partitioning Problem is NP-hard and NP-hard to approximate

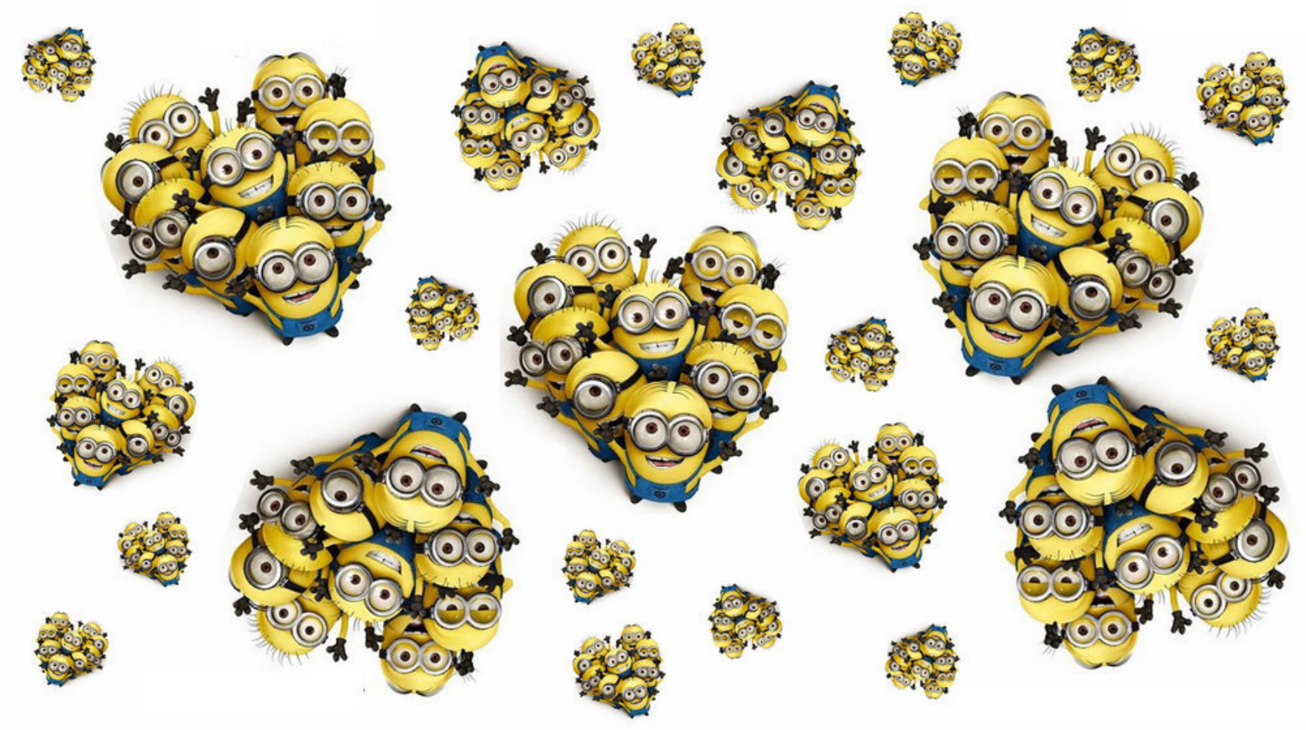
Solving Target Partitioning

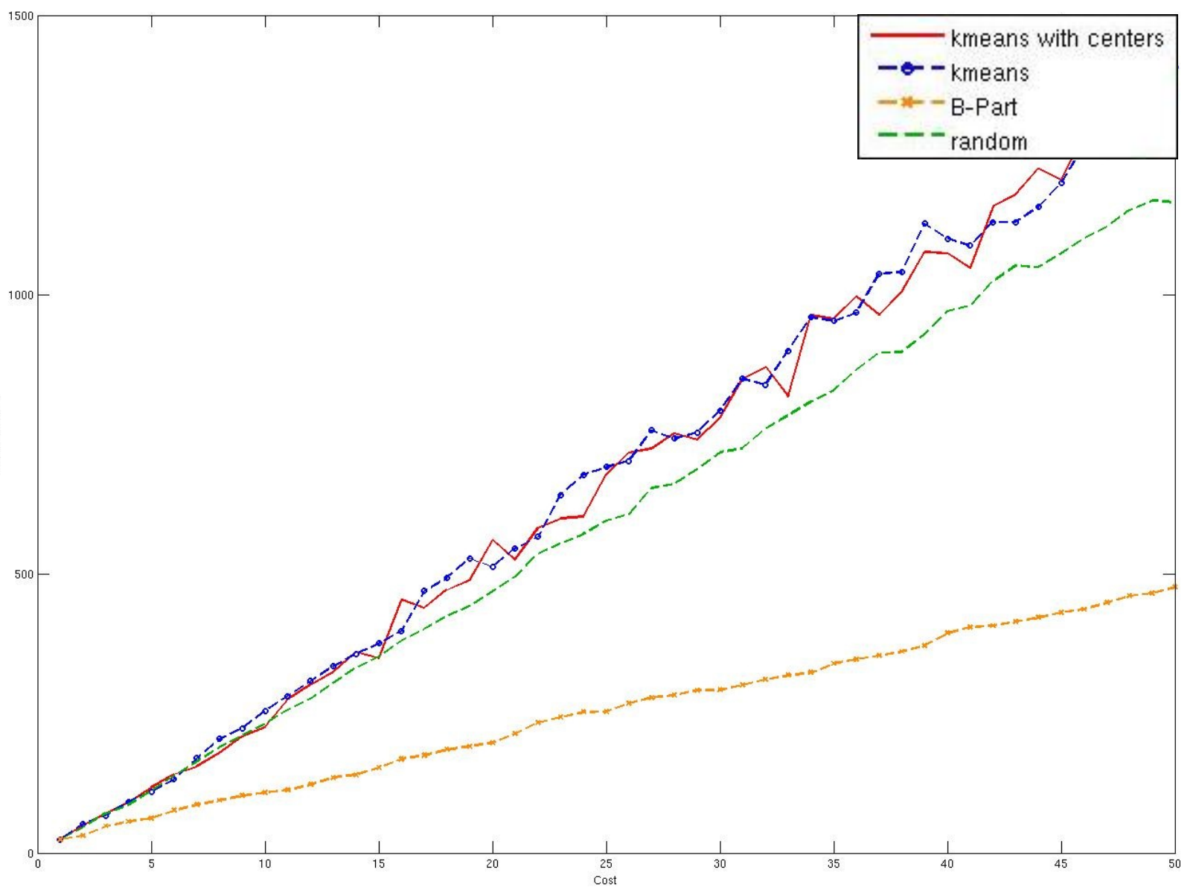
K-means

- Cluster points using K-means clustering to get cluster centers
- Find the matching between cluster centers and targets (min-weight perfect matching)

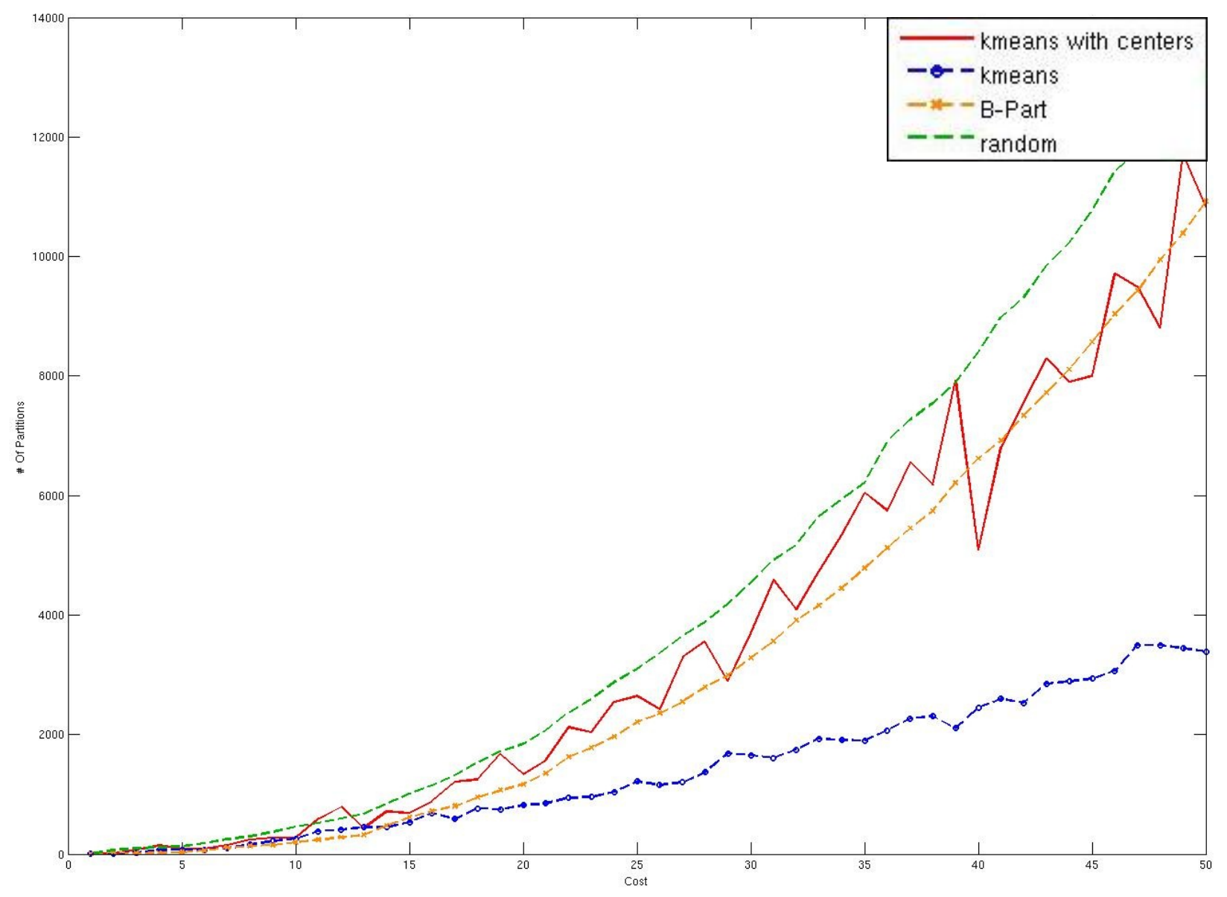
Benefit Partition

Assign each point to the partition which benefits the most from adding that point.





Random Data



Synthetic Data

K-means with Targets	K-means	Benefit Partitioning	Random partitioning
42422	30168	29187	35355

Intel Dataset

K-means with Targets	K-means	Benefit Partitioning	Random partitioning
4310291	4453967	95182	126192

PLOS Dataset

Target-ℓ Partitioning

Given $X = \{x_1, x_2, .., x_n\}$ and vectors $\tau_1, \tau_2, \ldots, \tau_k$ and ℓ, partition $X \setminus X_\ell$ into k groups such that $\Sigma \| \mu(C_i) - \tau_i \|_{(i=1..k)}$ is minimized.

Target-ℓ Partitioning Problem is NP-hard and NP-hard to approximate

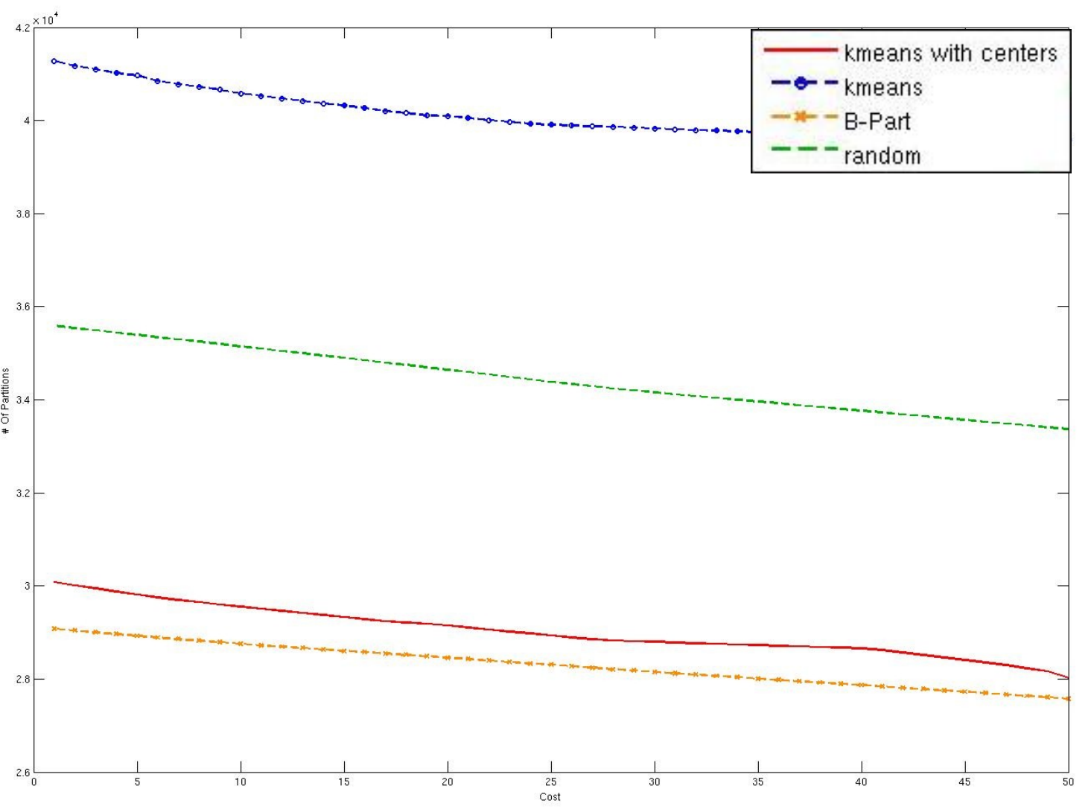
Solving Target-ℓ Partitioning

Step 1: Partition the points into k groups, using the algorithm for target partitioning algorithm.

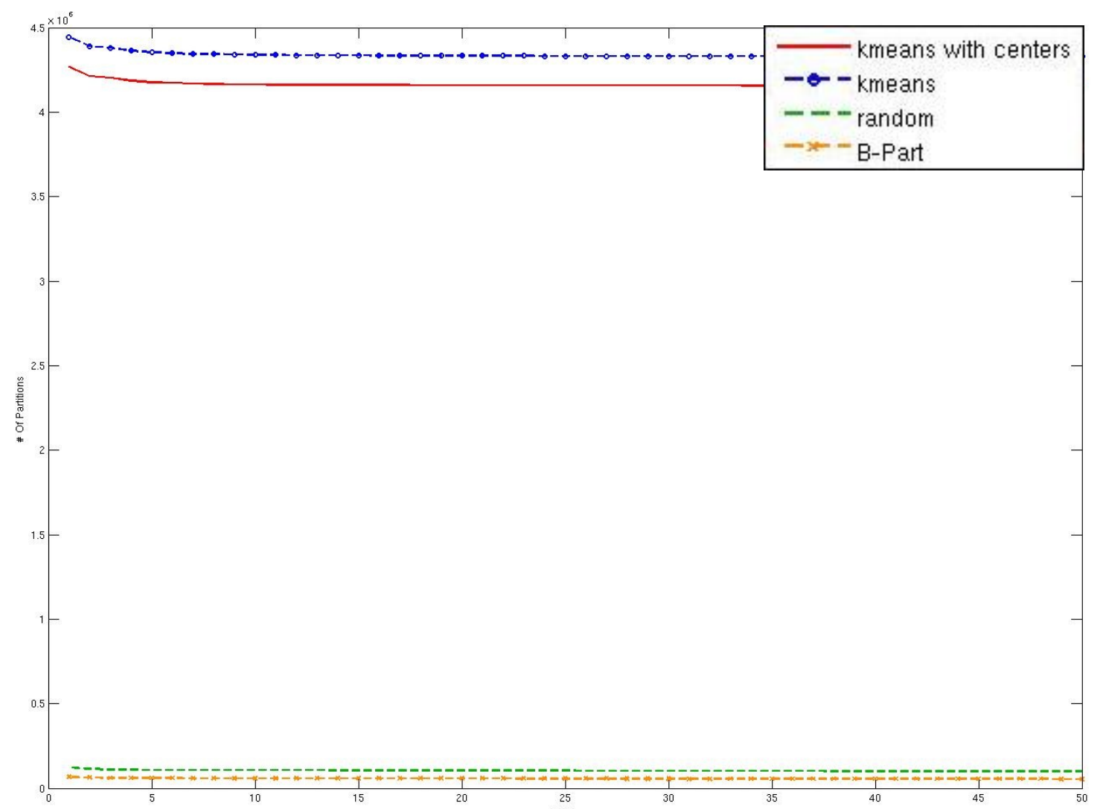
Step2: For each partitions find i= 1.. ℓ points to remove from, using algorithm for ℓ-removal problem.

Step3: Remove ℓ points:

$$D(i, j) = \max \{ D(i - 1, j - q) + d(i, q)\} \qquad 0 <= q <= j$$



Intel Dataset



PLOS Dataset

Detected Outliers in PLOS:

- "AID Enzymatic Activity Is Inversely Proportional to the Size of Cytosine C5 Orbital Cloud"

Twitted: 0	AVG: 38.55
HTML Page downloaded: 1583	AVG: 1734.49
PDF Downloaded: 262	AVG: 350.81
XML Downloaded: 150	AVG: 187.95

- "Performance of the PointCare NOW System for CD4 Counting in HIV Patients Based on Five Independent Evaluations"

Twitted: 23	AVG: 38.55
HTML Page downloaded: 704	AVG: 1734.49
PDF Downloaded: 71	AVG: 350.81
XML Downloaded: 9	AVG: 187.95

Reference

Selecting a Set of Characteristic Reviews, T. Lappas, M. Crovella, E. Terzi. ACM SIGKDD 2012