CURE: An Efficient Clustering Algorithm for Large Databases

Sudipto Guha
Stanford University

Rajeev Rastogi
Bell Laboratories

Kyuseok Shim
Bell Laboratories

Presented by Afnan Ahmad

Monday, November 20, 2017

YORK UNIVERSITÉ UNIVERSITY

Overview of the Paper

- Introduction
  - Drawbacks of Traditional Clustering Algorithms
  - Contributions of CURE
- CURE Algorithm
  - Hierarchical Clustering Algorithm
  - Random Sampling
  - Partitioning for Speedup
  - Labeling Data on Disk
  - Handling Outliers
- Experimental Results

YORK UNIVERSITÉ UNIVERSITY

Drawbacks of Traditional Clustering Algorithms

- PARTITIONING CLUSTERING
  - This category of clustering method try to reduce the data set into k clusters based on some criterion functions.
  - The most common criterion is square-error criterion.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon C_i} ||p - m_i||^2$$

  - This method favor to clusters with data points as compact and separated as possible

YORK UNIVERSITÉ UNIVERSITY

3

Drawbacks of Traditional Clustering Algorithms (cont)

- You may find error in case the square-error is reduced by splitting some large cluster to favor some other group.

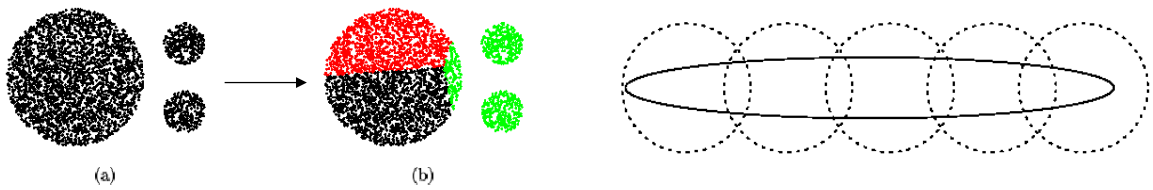


Figure: Splitting occur in large cluster by partitional method

YORK UNIVERSITÉ UNIVERSITY

4

Drawbacks of Traditional Clustering Algorithms (cont)

- Hierarchical Clustering
  - This category of clustering method try to merge sequences of disjoint clusters into the target k clusters base on the minimum distance between two clusters.
  - The distance between clusters can be measured as:
    - Distance between mean:

$$d_{mean}(C_i, C_j) = \| m_i - m_j \|$$

    - Distance between average point

$$d_{ave}(C_i, C_j) = {}^{1}/_{(n_i n_j)} \sum_{p \epsilon C_i} \sum_{p' \in C_i} \| p - p' \|$$

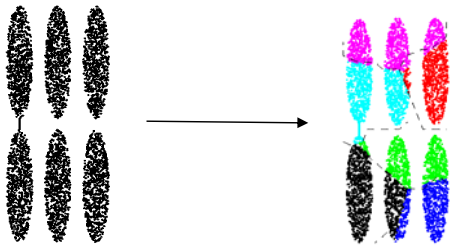
    - Distance between two nearest point within cluster

$$d_{min}(C_i C_j) = \min_{p \in C_i, p' \in C_j} \| p - p' \|$$

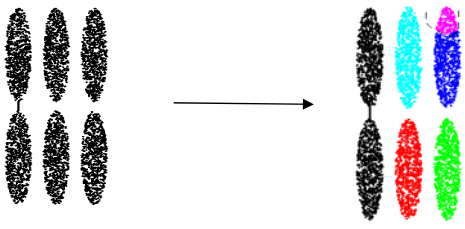
YORK UNIVERSITÉ UNIVERSITY

Drawbacks of Traditional Clustering Algorithms (cont)

- This method favor hyper-spherical shape and uniform data.
- Let's take some prolonged data as example:



Result of d_{mean}



Result of d_{min}

YORK UNIVERSITÉ UNIVERSITY

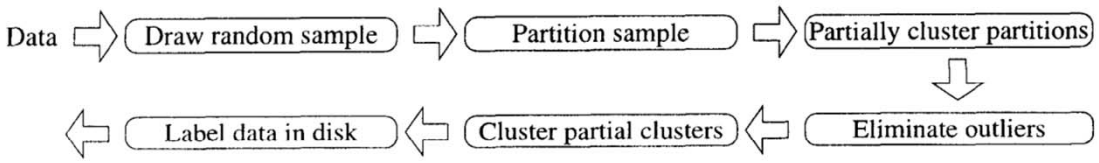
Contributions of CURE

- CURE can identify both spherical and non-spherical clusters.
  - It chooses a number of well scattered points as *representatives* of the cluster instead of one point - centroid.
- CURE uses *random sampling* and *partitioning* to speed up clustering.

YORK UNIVERSITÉ UNIVERSITY

7

Overview of CURE



YORK UNIVERSITÉ UNIVERSITY

8

CURE's Advantages

- More accurate:
  - Adjusts well to geometry of non-spherical shapes.
  - Scales to large datasets
  - Less sensitive to outliers
- More efficient:
  - Space complexity: $O(n)$
  - Time complexity: $O(n^2 \log n)$ ($O(n^2)$ if dimensionality of data points is small)

YORK UNIVERSITÉ UNIVERSITY

CURE Algorithm: Hierarchical Clustering Algorithm

- For each cluster, c well scattered points within the cluster are chosen , and then shrinking them toward the mean of the cluster by a fraction α.
- The distance between two clusters is then the distance between the closest pair of representative points from each cluster.
- The c representative points attempt to capture the physical shape and geometry of the cluster. Shrinking the scattered points toward the mean gets rid of surface abnormalities and mitigates the effects of outliers.

YORK UNIVERSITÉ UNIVERSITY

CURE Algorithm: Random Sampling

- In order to handle large data sets, *random sampling* is used to reduce the size of the input to CURE's clustering algorithm.
- [Vit85] provides efficient algorithms for drawing a sample randomly in one pass and using constant space.
- Although random sampling does have tradeoff between accuracy and efficiency, experiments show that for most of the data sets, with moderate sized random samples, very good clusters can obtained.

YORK UNIVERSITÉ UNIVERSITY

CURE Algorithm: Partitioning for Speedup

- In the first pass:
  - Partition the sample space into p partitions, each of size n/p.
  - Partially cluster each partition until the final number of clusters in each partition reduces to $n/pq, q > 1$.
- The advantages are
  - Reduce execution time
  - Reduce the input size and ensure it fits in main-memory by storing only the representative points for each cluster as input to the clustering algorithm.

YORK UNIVERSITÉ UNIVERSITY

CURE Algorithm: Labeling Data on Disk

- Input is randomly selected sample
- Have to assign the appropriate cluster labels to the remaining data points
- Each data point is assigned to the cluster containing the representative point closest to it.
- Advantage: using multiple points enables CURE to correctly distribute data points when clusters are non-spherical or non-union.

13

YORK UNIVERSITÉ UNIVERSITY

CURE Algorithm: Handling Outliers

- The number of points in a collection of outliers is typically much less than the number in a cluster.
- In the first phase, proceed with the clustering until # clusters decreases to $1/3$, then classify clusters with very few points (e.g., 1 or 2) as outliers.
- The second phase occurs toward the end. Small groups (outliers) are easy to be identified and eliminated.

14

YORK UNIVERSITÉ UNIVERSITY

Experimental Results - Algorithms

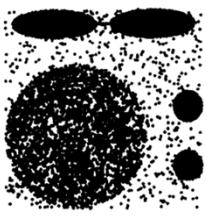
- BIRCH
- CURE
  - The partitioning constant $q = 3$
  - Two phase outlier handling
  - Random sample size = 2.5% of the initial data set size
- MST (minimum spanning tree)
  - When shrink factor = 0, CURE reduces to MST.

YORK UNIVERSITÉ UNIVERSITY

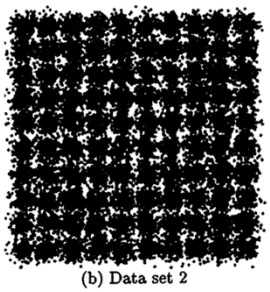
Experimental Results – Data Sets

Experiment with data sets of two dimensions

- Data set 1 contains one big and two small circles.
- Data set 2 consists of 100 clusters with centers arranged in a grid pattern and data points in each cluster following a normal distribution with mean at the cluster center.



(a) Data set 1

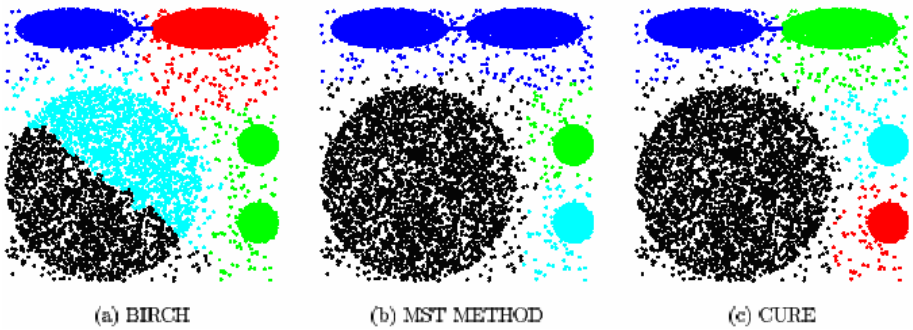


(b) Data set 2

YORK UNIVERSITÉ UNIVERSITY

Experimental Results – Quality of Clustering

- BIRCH cannot distinguish between the big and small clusters.
- MST merges the two ellipsoids.
- CURE successfully discovers the clusters in Data set 1.



(a) BIRCH (b) MST METHOD (c) CURE

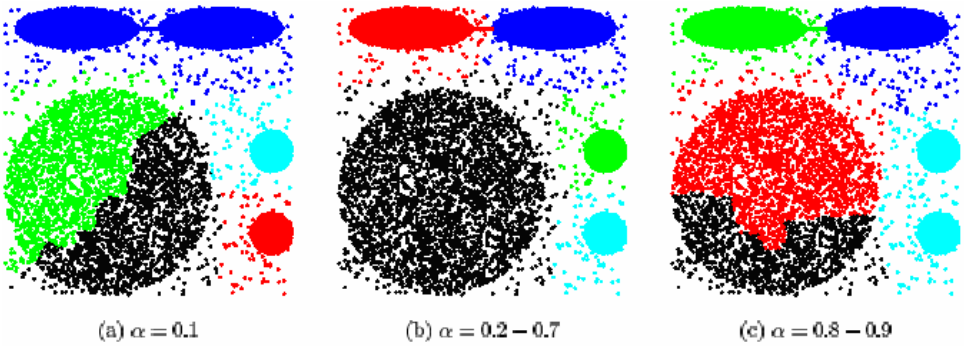
17

YORK UNIVERSITÉ UNIVERSITY

Experimental Results – Sensitivity to Parameters

Shrink Factor a:

- 0.2 – 0.7 is a good range of values for a.



(a) $\alpha = 0.1$ (b) $\alpha = 0.2 - 0.7$ (c) $\alpha = 0.8 - 0.9$

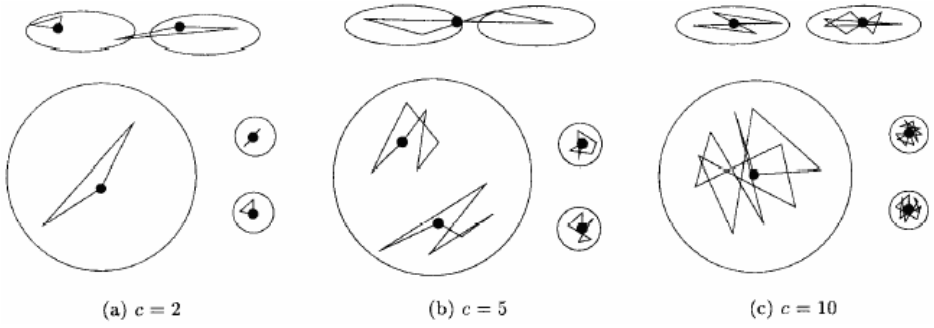
18

YORK UNIVERSITÉ UNIVERSITY

Experimental Results – Sensitivity to Parameters (Cont)

Number of Representative Points c:

- For smaller values of c, the quality of clustering suffered.
- However, for values of c greater than 10, CURE always found right clusters.

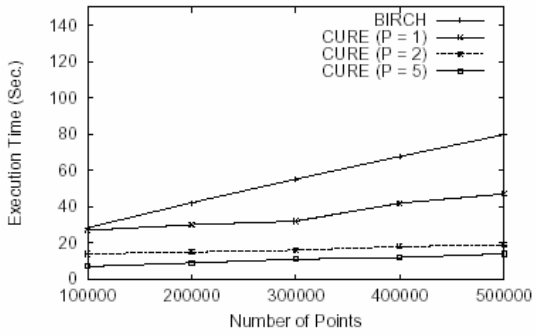


19

Experimental Results – Comparison of Execution time to BIRCH

Run both BIRCH and CURE on Data set 2:

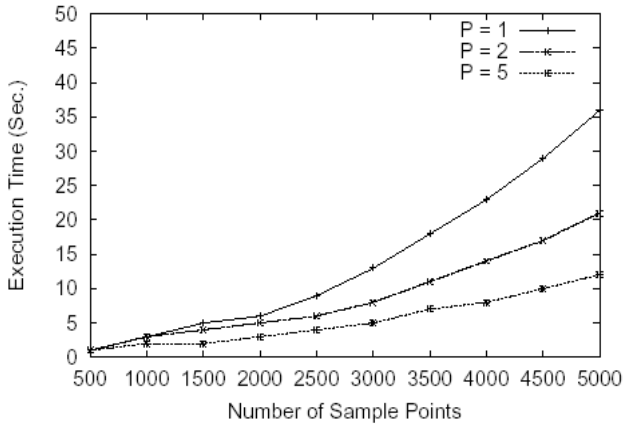
- CURE execution time is always lower than BIRCH
- Partitioning improves CURE's running time > 50%
- As sample size goes up, CURE's execution time only slightly increases due to fixed sample size



20

Experimental Results – Scale up experiment

Run CURE on Data set 1:



21

Conclusion

- CURE can handle large databases efficiently.
- CURE can effectively detect proper shape of the cluster with the help of scattered representative point and centroid shrinking.
- CURE can reduce computation time and memory loading with random sampling and 2 pass clustering
- CURE can effectively remove outlier.
- The quality and effectiveness of CURE can be tuned be varying different ***s,p,c,*** α to adapt different input data set.

22

Thank You!

Any Questions?

YORK UNIVERSITÉ UNIVERSITY