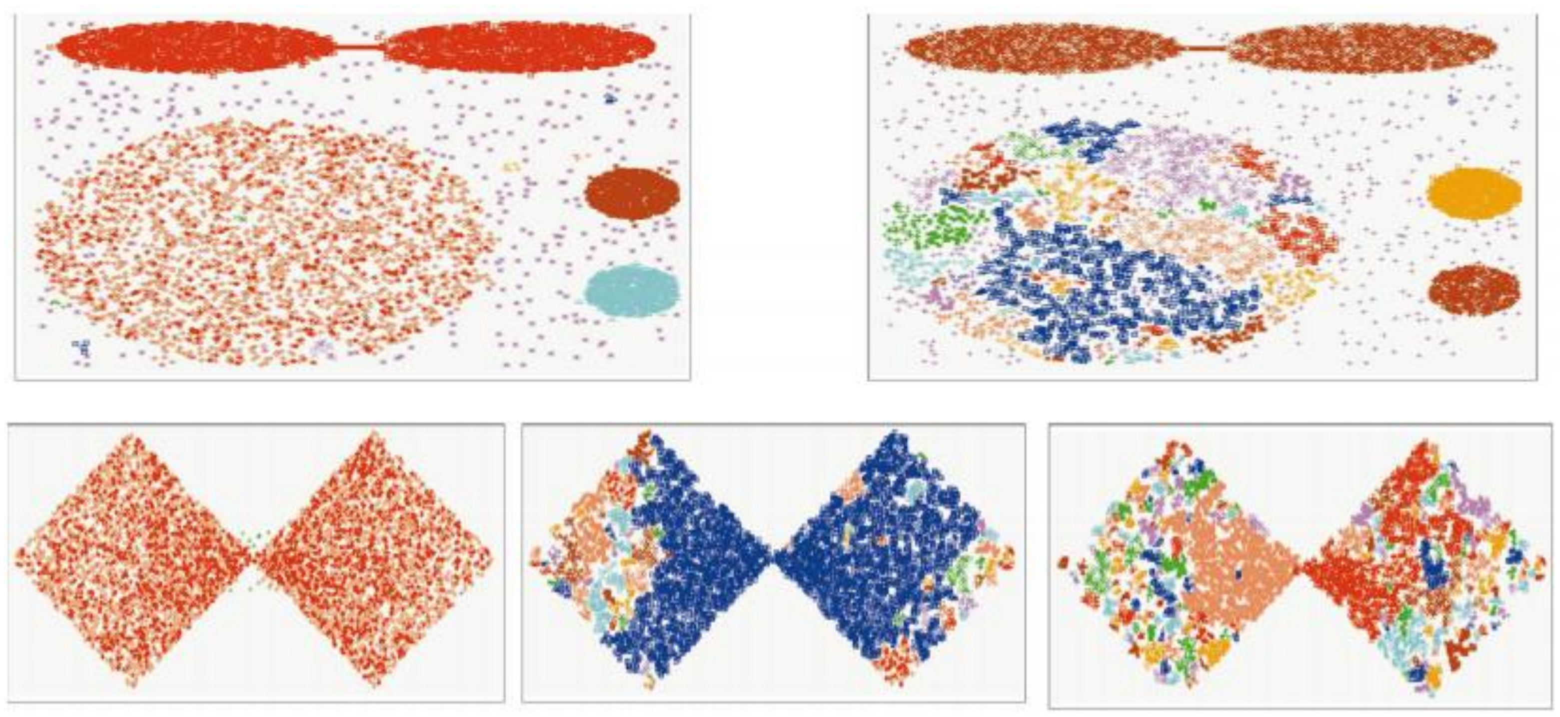
Improving PreDeCon With Graph Based Approach

Shahiduz Zaman (0905091) and Md. Momen Bhuiyan (0905099)

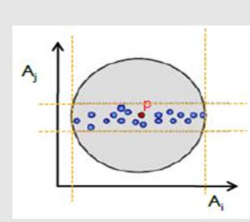
1. **Problem Definition:** Subspace clustering is fairly efficient to cluster high dimensional data. PreDeCon [1] is a density based subspace clustering variant of DBSCAN [2]. Two problems inherent in PreDeCon is:
   I. It can't differentiate between two close clusters.
   II. It is very sensitive to parameter.



2. **Existing Work:**

- PreDeCon extends DBSCAN to high dimension spaces by incorporating the notion of dimension preferences in the distance function
- For each point p, it defines its subspace preference vector:

$$\overline{W} = (w_1, w_2, \ldots, w_d) \qquad w_i = \begin{cases} 1 & \text{if } VAR_i > \delta \\ k & \text{if } VAR_i \le \delta \end{cases}$$



- VAR_{Ai} is the variance along dimension i in $N_\varepsilon(p)$:

$$VAR_{A_i}(N_\varepsilon(p)) = \frac{\sum_{q \in N_\varepsilon(p)} \left(dist(\pi_{A_i}(p), \pi_{A_i}(q)) \right)^2}{|N_\varepsilon(p)|}$$

δ, k(k>>1) are input parameters

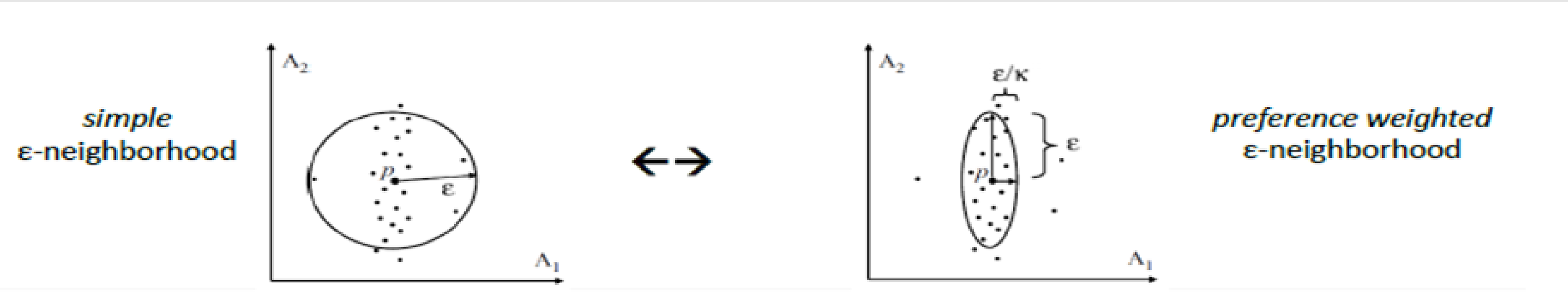
- Preference weighted distance function:

$$dist_p(p,q) = \sqrt{\sum_{i=1}^{d} \frac{1}{w_i} (\pi_{A_i}(p) - \pi_{A_i}(q))^2}$$

$$dist_{pref}(p,q) = \max\{dist_p(p,q), dist_q(q,p)\}$$

- Preference weighted ε-neighborhood:

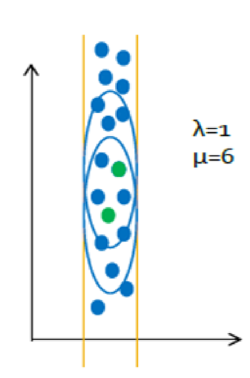
$$N_\varepsilon^{\overline{w_p}}(p) = \{x \in D \mid dist_{pref}(p,x) \le \varepsilon\}$$



- Preference weighted core points:

$$CORE_{den}^{pref}(p) \Leftrightarrow PDIM(N_\varepsilon(p)) \le \lambda \wedge \left| N_\varepsilon^{\overline{w_0}}(p) \right| \ge \mu$$

- Direct density reachability, reachability and connectivity are defined based on preference weighted core points.
- A *subspace preference cluster* is a maximal density connected set of points associated with a certain subspace preference vector.



3. **Solution for Problem I:** Simplest way to solve the 1st problem is to convert it into a problem of graph and apply a graph cut algorithm to each cluster from PreDeCon.
   a. **Converting To Graph:** We considered two ways to apply connectivity to a data point.
      i. **k-Nearest Neighbors:** Requires parameter k.
      ii. **Direct Density Reachable:** Requires nothing new.
   b. **Graph Cut Algorithm:** Two considerations were taken here:
      i. **Min-Cut Algorithm:** Karger's algorithm [3] finds all min-cut with complexity of $O(n^2 \log^3 n)$ with error probability of O(1/n).
      ii. **Sparsest Cut Algorithm:** Although sparsest cut is a NP-Hard there is an approximation with complexity of O(√logn) [4].

4. **Experimental Result:** We used ELKI to implement direct density reachability and min-cut algorithm. Then PreDeCon and our Improved version is applied on data similar to below and evaluation result shows significant improvement.

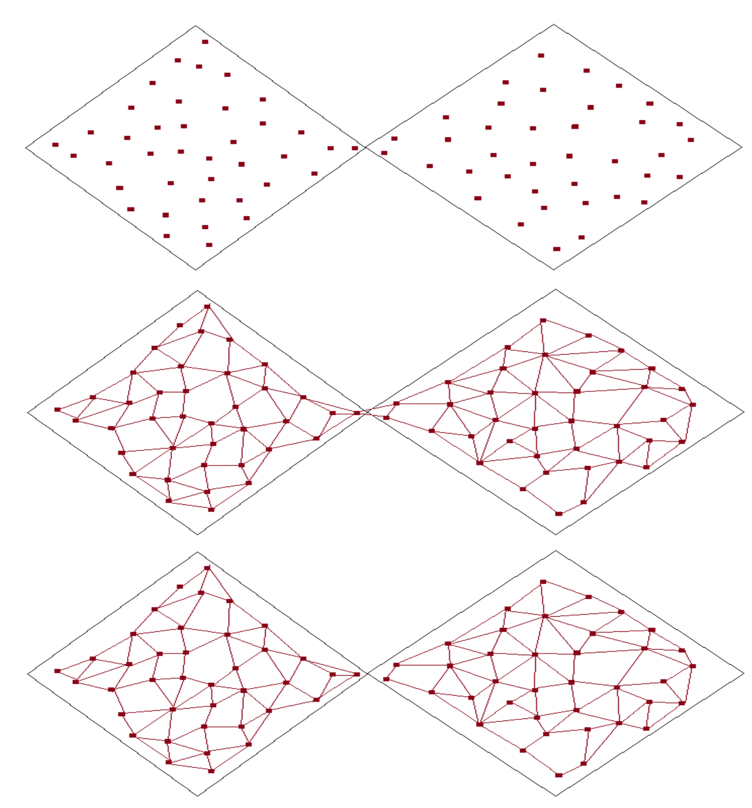


Fig 1. Dataset 1

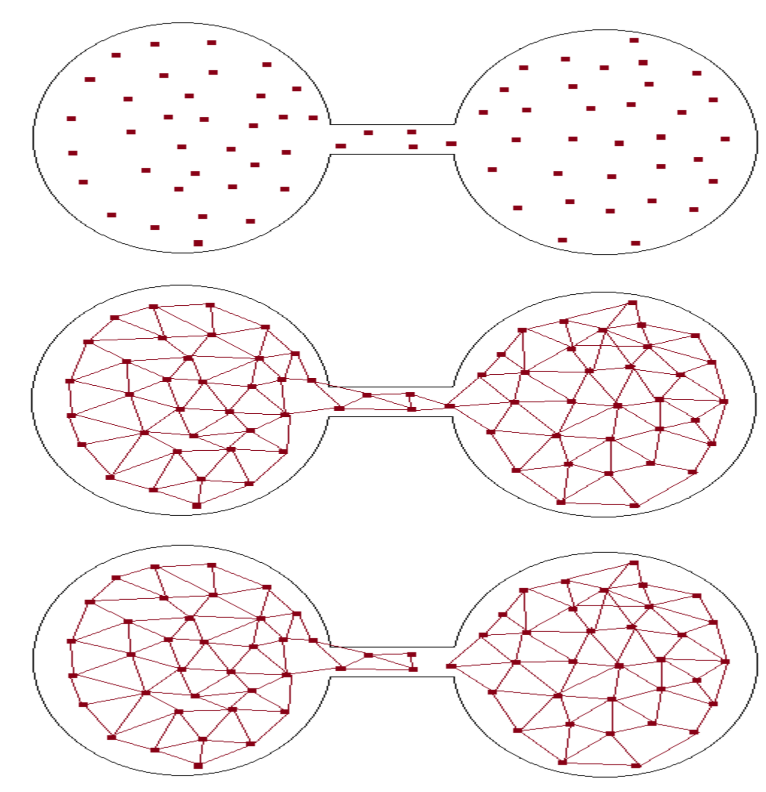


Fig 2. Dataset 2

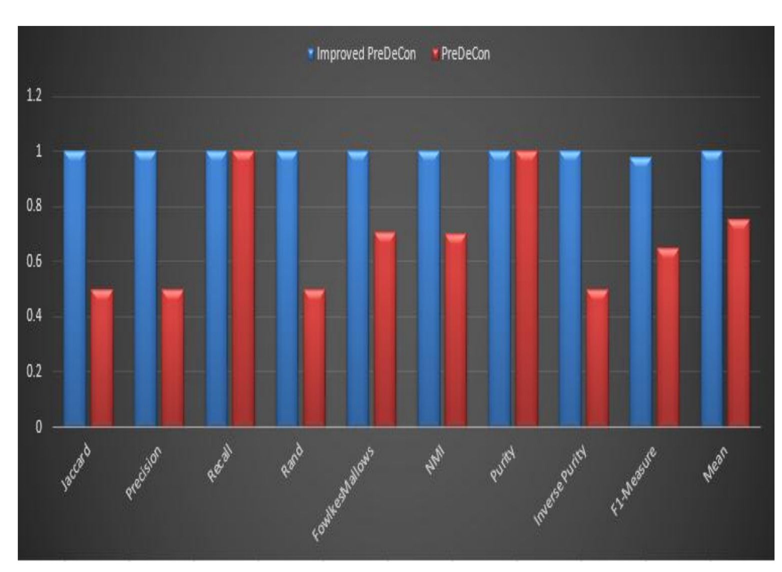


Fig 3. Evaluation Result for Dataset 1

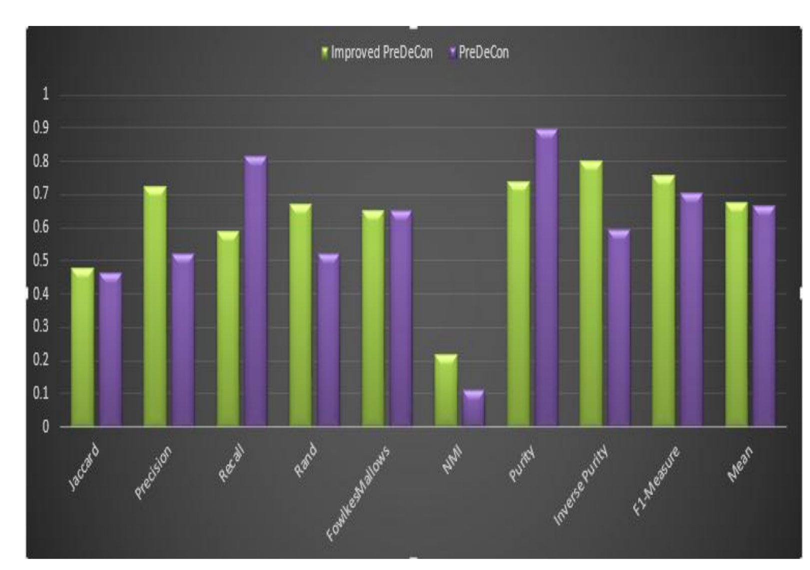


Fig 4. Evaluation Result for Dataset 2

4. **Proposed Solution for Problem II:** For the 2nd Problem we are proposing to use relative variance in a dimension to select core-point in the algorithm along with MinPts. Here is a suggestion:

$$CORE_{den}^{pref}(p) \Leftrightarrow PDIM(N_\varepsilon(p)) \le \lambda \wedge \left| N_\varepsilon^{\overline{w_0}}(p) \right| \ge \mu$$

Or

$$CORE_{den}^{pref}(p) \Leftrightarrow \frac{V_{AR_{A_i}}(N_\varepsilon(p))}{V_{AR_{A_i}}()} > \tau$$

5. **Conclusion:** Our approach to solve the 1st problem is independent of the shape of the cluster. Our future work includes to implement sparsest cut algorithm in ELKI and experiment on the 2nd problem without increasing complexity of the clustering.

6. **Reference:**

[1] Bohm, Christian, K. Railing, H-P. Kriegel, and Peer Kroger. "Density connected clustering with local subspace preferences." In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pp. 27-34. IEEE, 2004.

[2] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, vol. 96, no. 34, pp. 226-231. 1996.

[3] Karger, David R. "Global Min-cuts in RNC, and Other Ramifications of a Simple Min-Cut Algorithm." In *SODA*, vol. 93, pp. 21-30. 1993.

[4] Arora, Sanjeev, Satish Rao, and Umesh Vazirani. "Expander flows, geometric embeddings and graph partitioning." *Journal of the ACM (JACM)* 56, no. 2 (2009): 5.

Department of Computer Science and Engineering (CSE), BUET