Can DBSCAN Be Improved by Robust Preprocessing?

Jan Thiele

CTU in Prague, FJFI

Overview

1. Data whitening
2. Geometric median
3. Pursuit method
4. DBSCAN
5. Results

Data whitening

- data normalization
- $\mathrm{E}X = \mathbf{0}$ and $\mathrm{var}X = \mathbf{I}$

$$\mathbf{W} = \mathbf{X}_{\mathrm{S}}^{\mathrm{T}} \left(\frac{\mathbf{u}_{(1)}}{\lambda_{(1)}}, \frac{\mathbf{u}_{(2)}}{\lambda_{(2)}}, \dots \frac{\mathbf{u}_{(D)}}{\lambda_{(D)}} \right) \in \mathbf{R}^{n \times D}.$$

Geometric median

1. statistically robust alternative to center of mass
2. $\mathbf{y} \in \mathbb{R}^n$,where

$$\sum_{i=1}^{m} ||\mathbf{x}_i - \mathbf{y}||_2 = \min .$$

3. Weiszfeld algorithm [4]

Pursuit method

1. data decorrelation
2. optimal linear combination of base vectors
3. robust variance

$$||\mathbf{w}_1^\star|| = \min,$$

$$\mathrm{var}^\star(\mathbf{z}) = 1.$$

$$\frac{||\mathbf{w}_1^\star||^2}{\mathrm{var}^\star(\mathbf{z})} = \min_{\mathbf{w}_1^\star \neq \mathbf{0}},$$

Robust Standard Deviation Estimates

- S_n [1]

$$S_n = 1.1926 \cdot \operatorname*{lomed}_{i=1,\ldots,n} \cdot \operatorname*{himed}_{j=1,\ldots,n} |x_i - x_j|$$

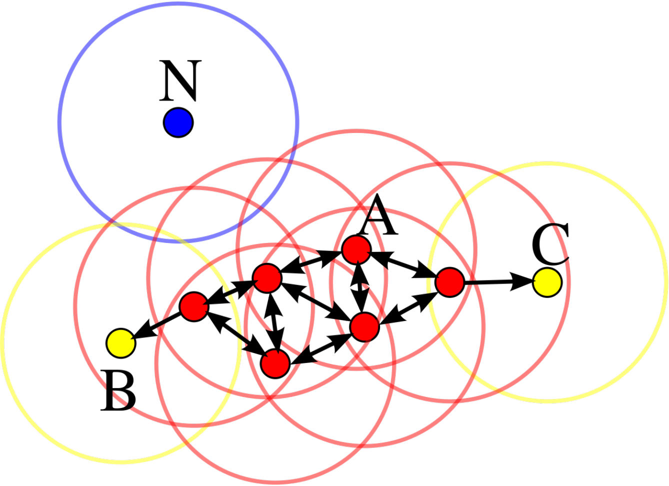
- Q_n [1]

$$Q_n = 2.2219 \cdot \{|x_i - x_j|; i < j\}_{(\lfloor \frac{n}{4} \rceil)}$$

- MAD_n [3]

$$MAD_n = 1.4826 \cdot \mathrm{med}_i |x_i - \mathrm{med}_j x_j|$$

DBSCAN



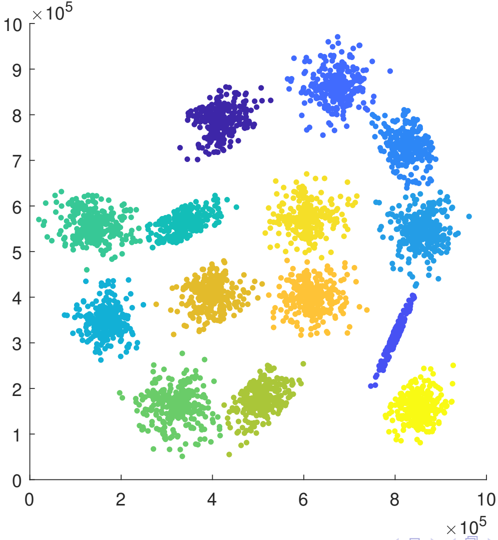
1

[1]By Chire - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=17045963

- minpts - minimal number of neighbours
- ε - maximal distance to neighbours
- core points
- border points
- noise points (outliers)

Results

Dataset [2]



Non-robust approach

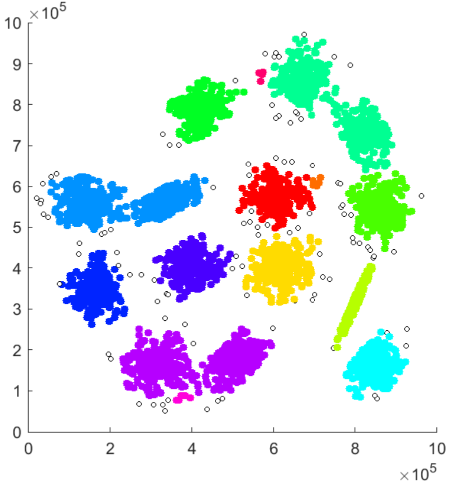


Figure: minpts = 4, $\varepsilon = 0.95$, 2.2 % outliers

Robust approach

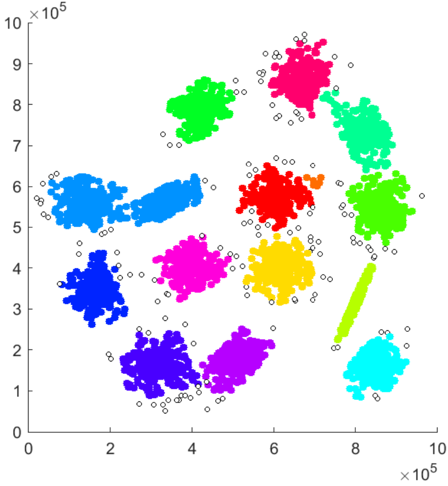


Figure: minpts = 5, $\varepsilon = 0.8$, 1.8 % outliers

References

C. Croux and P. Rousseeuw.

Time-efficient algorithms for two highly robust estimators of scale.

Computational Statistics, Vol. 1, 1:411–428, 01 1992.

M. Liu, B. Liu, C. Zhang, W. Wang, and W. Sun.

Semi-supervised low rank kernel learning algorithm via extreme learning machine.

International Journal of Machine Learning and Cybernetics, 8, 06 2017.

P. J. Rousseeuw and C. Croux.

Alternatives to the median absolute deviation.

Journal of the American Statistical Association, 88(424):1273–1283, 1993.

E. Weiszfeld and F. Plastria.

On the point for which the sum of the distances to n given points is minimum.

Annals of Operations Research, 167(1):7–41, Mar 2009.