CSC4008 Project: Improvement of DBSCAN using Rank Minimization

WEN Zhenduo, LI Zhuoyu

Chinese University of Hong Kong, Shenzhen

May 11, 2021

Outline

1. DBSCAN
2. Robustness achieved by Rank Minimization
    - Sparsity and Low-Rank Property of Adjacency Matrix
    - NP Hardness Optimization Problem
    - Analytic Solution under fully-observed Condition
3. Implementation
    - Data Cleaning
    - Feature Selection
    - Result
4. Limitations

Section 1

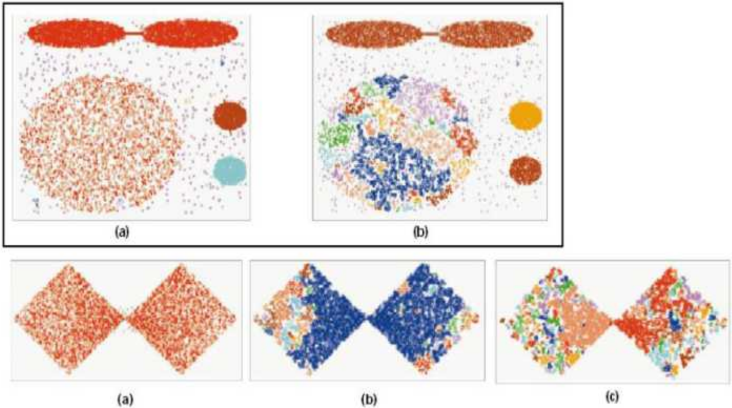
DBSCAN

Density Based Spatial Clustering of Applications with Noise

- start with arbitrary point p
- merge points that are located in preset distance *Eps*
- if p is the core point (reaches minimal number surrounding points within *Eps*), we find a cluster; otherwise continue the iteration until all points are processed.

Pros & Cons

- robust to noise
- can identify any shapes of clusters
- **Highly sensitive to preset values.**



Section 2

Robustness achieved by Rank Minimization

Sparsity and Low-Rank Property of Adjacency Matrix

If the dataset is well-behaved(if there are possible clustering patterns), the Adjacency Matrix returned by our clustering algorithm has the following two properties:

- **Sparsity**
- **Low-Rank**

This provides the possibility of improving DBSCAN by applying rank minimization on the adjacency matrix.

Sparsity & Low-Rank: Intuitive Explanation

Simple Example:

Consider we have five elements $\{1, 2, 3, 4, 5\}$. They are clustered into 3 sets: $\{1, 3\}$, $\{2\}$, $\{4, 5\}$. The Adjacency Matrix is:

$$S = \begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

Theoretically, the rank of Adjacency Matrix equals the amount of clusters.

Approximation

NP-hard rank minimization

$$\begin{aligned} &\text{minimize} \quad rank(X) \\ &\text{subject to} \quad X_{ij} = M_{ij},\ (i,j) \in \Omega \end{aligned}$$

Tightest Convex Relaxation of the above problem

Provided that the number of samples obey $m \geq Cn^{6/5} r \log(n)$, where n follows $M \in R^{n \times n}$, m is the available sample entries, r is the rank of M:

$$\begin{aligned} &\text{minimize} \quad ||X||_* \\ &\text{subject to} \quad X_{ij} = M_{ij},\ (i,j) \in \Omega \\ &\textit{where } ||X||_* \text{ is the nuclear norm of } X. \\ &\Omega \text{ is the observable set} \end{aligned}$$

Analytic Solution under fully-observed condition

Tightest Convex Relaxation of the above problem

$$
\begin{aligned}
\text{minimize} \quad & ||X||_* \\
\text{subject to} \quad & X_{ij} = M_{ij},\ (i,j) \in \Omega
\end{aligned}
$$

That is, to solve:

$$\arg\min_X \left\{ \frac{1}{2} ||P_\Omega(X) - P_\Omega(M)||_F^2 + \tau ||X||_* \right\}$$

where $P_\Omega(*)$ is the operator to project matrices onto the observable set Ω: $P_\Omega(X_{ij}) = X_{ij}$ if $(i,j) \in \Omega$.

In general, the above optimization problem is non-convex. We need to carefully design an iteration and warm-restart using sequences of τ.

Analytic Solution under fully-observed condition

But in terms of DBSCAN, the algorithm tries the distance between pairs of data points and returns the cluster result. Hence, the adjacency matrix is **fully-observed**. We do not need the projection operator P_Ω. We claim that, the optimization problem then has an analytic solution:

Solution given by Singular Value Shrinkage operator

$$D_\tau(M) = \arg\min_X \left\{ \frac{1}{2}||X - M||_F^2 + \tau||X||_* \right\}$$

where

$$\begin{aligned} &\tau \text{ is given} \\ &M = U \cdot \Sigma \cdot V' \\ &D_\tau(M) = U \cdot [(\sigma_1 - \tau)_+, ..., (\sigma_m - \tau)_+] \cdot V' \end{aligned}$$

Analytic Solution under fully-observed condition

Proof

The function $h_{\tau,M}(X) = \tau||X||_* + \frac{1}{2}||X - M||_F^2$ is convex, so there exists a unique minimizer. We need to prove it equals $\hat{X} = D_\tau(M)$.
Note that $\hat{X}$ is the unique minimizer if and only if subgradient of $h_{\tau,M}(\hat{X})$ contains zero. That is:

$$0 \in \hat{X} - M + \tau\partial||\hat{X}||_*$$

where $\tau\partial||\hat{X}||_*$ is the set of subgradients of nuclear norm. We choose the following subgradients, suppose $X = U\Sigma V'$:

$$\partial||\hat{X}||_* = \left\{UV' + W : U'W = 0, WV = 0, ||W||_2 \leq 1\right\}$$

Analytic Solution under fully-observed condition

Proof

Decompose M into 2 parts:

$$M = U_0 \Sigma_0 V_0' + U_1 \Sigma_1 V_1'$$

where U_0, V_0 contains singular vectors associated with singular values greater than τ, thus

$$\hat{X} = U_0(\Sigma_0 - \tau I)V_0'$$

Therefore

$$M - \hat{X} = \tau(U_0 V_0' + W), \quad W = \tau^{-1} U_1 \Sigma_1 V_1'$$

with $U_0' W = 0, WV_0 = 0, ||W||_2 \leq 1$. W is the subgradient of $||X||_*$. Thus $M - \hat{X} \in \tau \partial ||\hat{X}||_*$, QED.

Section 3

Implementation

Missing 60 minutes

- The data for 2016-04-30 is incomplete. Thus is abandoned.
- Most users then have 172740 records, which magically missing 60 minutes, with respect of 172800 minutes between 2016-01-01 00:00:00 and 2016-04-29 23:59:00

Missing 60 minutes

dataset 7	
3273	172740
3268	172740
2931	172740
2945	172740
2953	172740
2965	172740
2980	172740
2986	172740
3036	172740
3039	172740
3104	172740
3126	172740
3134	172740
3192	172740
3221	172740
2925	172740
3009	172680
3044	172620
3092	86700

Figure: Records in dataset 7

Daylight saving time

```
3943710,2016-03-13 01:59:00-06,4213,0.602
3943711,2016-03-13 01:59:00-06,4505,0.701
3943712,2016-03-13 01:59:00-06,4313,0.777
3943713,2016-03-13 01:59:00-06,4601,1.123
3943714,2016-03-13 01:59:00-06,4946,0.5429999999999999
3943715,2016-03-13 01:59:00-06,4703,0.41
3943716,2016-03-13 01:59:00-06,4447,0.7809999999999999
3943717,2016-03-13 03:00:00-05,4957,1.44
3943718,2016-03-13 03:00:00-05,4031,0.276
3943719,2016-03-13 03:00:00-05,4193,0.33399999999999996
3943720,2016-03-13 03:00:00-05,4590,0.365
3943721,2016-03-13 03:00:00-05,4944,0.21100000000000002
3943722,2016-03-13 03:00:00-05,4298,0.26899999999999996
3943723,2016-03-13 03:00:00-05,4830,0.0
3943724,2016-03-13 03:00:00-05,4220,0.2
```

Figure: Records on 2016-03-13 Midnight

Daylight saving time (DST) is to set clocks forward by one hour in the spring ("spring forward") and set clocks back by one hour in autumn ("fall back") to return to standard time.

Data Cleaning

- Set all negative electricity usage to 0
- Eliminate users that has significantly less amount of records.
- Use ARIMA model to fit the missing values.
- Take the average consumption of each user in a particular minute in a day.(1 minute consumption by taking average of same minute on 120 days)

Feature Selection

MCI

User's MCI reflects the true cost of the user to the grid in a certain time period:

$$
\begin{aligned}
MCI_i &= \lim_{\Delta \to 0} \frac{C(L + \frac{\Delta L_i}{||L_i||_1}) - C(L)}{\Delta} \\
&= \lim_{\Delta \to 0} \frac{\sum_{t=1}^{T} (aL^t + b)\frac{\Delta l_i^t}{||l_i||_1} + \left(\frac{\Delta l_i^t}{||l_i||_1}\right)^2}{\Delta} \\
&= \sum_{t=1}^{T} (aL^t + b)\frac{l_i^t}{||l_i||_1}
\end{aligned}
$$

where $C(*)$ is the cost function of electricity consumption L with a quadratic form: $C(L) = \frac{1}{2}aL^2 + bL + c$ In our project, we take $a = 0.01, b = 20$.

Feature Selection

We choose the following two features:

- User's MCI in day: 7:00 to 19:00
- User's MCI in night: 19:01 to 6:59

Result

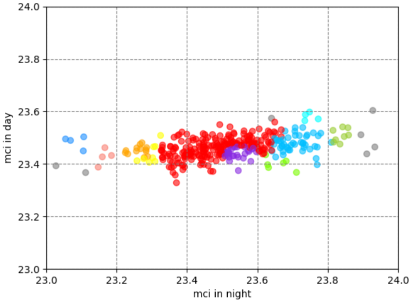


Figure: Clustering before Rank minimization

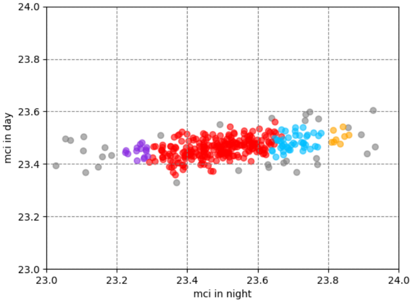
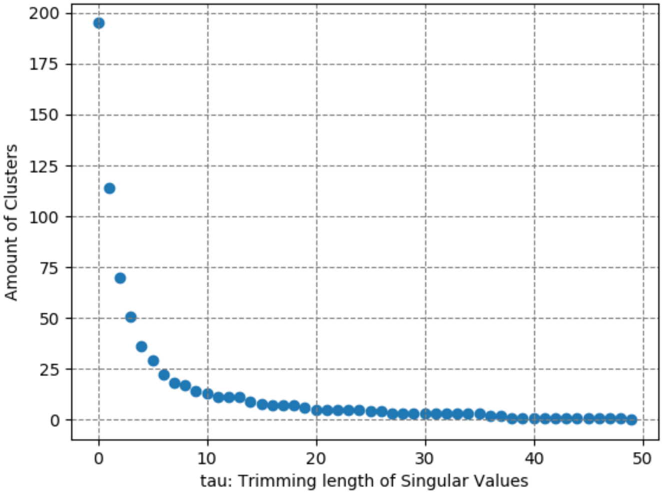


Figure: Clustering after rank minimization

Shrinking Singular Values and Amount of Clusters



Limitations

- The approximate optimization does not always reach the "best" solution. (thinking about l_1 minimization may not approximate l_0 minimization when there are several sparse vectors in null space)
- Perform not so good in moon-shape or round-shape distributed data.
- Still depends the performance of original algorithm. (It cannot improve a lot if the original cluster result is far from desirable.)

References

[1] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[2] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *The Journal of Machine Learning Research*, 15(1):2213–2238, 2014.

[3] Jingshi Cui, Haoxiang Wang, Chenye Wu, and Yang Yu. Robust data-driven profile-based pricing schemes. *arXiv preprint arXiv:1912.05731*, 2019.

[4] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schroedl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.

[5] Yang Yu, Guangyi Liu, Wendong Zhu, Fei Wang, Bin Shu, Kai Zhang, Nicolas Astier, and Ram Rajagopal. Good consumer or bad consumer: Economic information revealed from demand profiles. *IEEE Transactions on Smart Grid*, 9(3):2347–2358, 2017.

The End