Paper ID: TU2.PG.2

Session: TU2.PG: Adaptation and Learning over Graphs

Graph Metric Learning via Gershgorin Disc Alignment

Cheng Yang[1]

Gene Cheung[1]

Wei Hu[2]

[1]York University, Canada

[2]Peking University, China

May 5th 2020

Outline

- Background on metric learning
- Related works
- Contribution
- Preliminaries
- **Graph metric learning**
- Results

Background

$$\delta_{i,j}(\mathbf{M}) = (\mathbf{f}_i - \mathbf{f}_j)^\top \mathbf{M} (\mathbf{f}_i - \mathbf{f}_j)$$

$\mathbf{f}_i \in \mathbb{R}^K$
Feature vector for sample i

Metric matrix

- Mahalanobis distance [1].

- ***Metric learning***: find $\mathbf{M} \in \mathbb{R}^{K \times K}$ that minimizes a chosen objective function $Q(\mathbf{M})$ subject to $\mathbf{M} \succ 0$.

convex and *differentiable*

Positive definite (PD)

[1] P. C. Mahalanobis, "On the generalized distance in statistics," Proceedings of the National Institute of Sciences of India, vol. 2, no. 1, pp. 49–55, April 1936.

Related works

- PD cone: Gradient descent / projection[2].

Step size

$$\mathbf{M}^{t+1} := \mathrm{Pr}\left(\mathbf{M}^t - \alpha \nabla Q(\mathbf{M}^t)\right)$$

➢ ***computation-expensive projection***
1) eigen-decomposition of $\mathbf{M}$.
2) soft-thresholding of eigenvalues.

Proximal gradient

➢ optimizing diagonal entries only [3]

$$\begin{bmatrix} m_{1,1} & & & \\ & m_{2,2} & & 0 \\ & 0 & \ddots & \\ & & & m_{K,K} \end{bmatrix}$$

- Sparse / Low-rank based methods [4].
  ➢ simply excludes the full-rank $\mathbf{M}$ with only positive diagonal entries.

Degrade the ***metric quality*** due to ***restricted search spaces***.

- Why **full-rank**?
  ➢ incorporates the **diagonal-only [3]** case.

[2] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell, "Distance metric learning, with application to clustering with side-information," NIPS'02.
[3] J. Zhang and L. Zhang, "Efficient stochastic optimization for low-rank distance metric learning," in AAAI, Feb. 2017, pp. 933–939.
[4] C. Yang, G. Cheung, and V. Stankovic, "Alternating binary classifier and graph learning from partial labels," in APSIPA, Nov. 2018, pp. 1137–1140.

Contributions

- A metric learning framework.
  - 1) Projection-free.
  - 2) For a general $Q(\mathbf{M})$. → Convex and differentiable

➢ Step 1: Define $\mathbf{M} \in \mathcal{S}$. → *general graph Laplacian* matrices search space.

- **self-loops**: relative importance among K features.
- **edge weights**: pairwise feature correlations.

➢ Step 2: Rewrite the PD cone constraint $\mathbf{M} \succ 0$ as signal-adaptive ***linear constraints*** via Gershgorin disc alignment [5].

➢ Step 3: optimize $\mathbf{M}$ { 1) Diagonal terms. 2) Off-diagonal terms. } as LP's via *Frank-Wolfe* iterations.

ICASSP2020 Barcelona

[5] Y. Bai, F. Wang, G. Cheung, Y. Nakatsukasa, and W. Gao, "Fast graph sampling set selection using Gershgorin disc alignment," to appear *IEEE TSP*, 2020.

Preliminaries

- An undirected graph. $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{W}\}$

A node (**feature**) set of cardinality $|\mathcal{V}| = K$

weighted **adjacency** matrix

edge set

each edge $(i, j) \in \mathcal{E}$ has a weight $w_{i,j}$ → similarity between *i* and *j*

- Generalized graph Laplacian.

$$\mathbf{L}_g = \mathbf{D} - \mathbf{W} + \mathrm{diag}(\mathbf{W})$$

degree matrix $d_{i,i} = \sum_{j=1}^{K} w_{i,j}$

ICASSP2020 Barcelona

Graph metric learning

- Graph metric matrix $\mathbf{M}$.

generalized graph Laplacian

1) positive edge weights $m_{i,j} \leq 0, i \neq j$
2) positive node degrees $m_{i,i} > 0$
3) may have self-loops with $w_{i,i} > -\sum_{j \mid j \neq i} w_{i,j}$

- Irreducible graph [6]
  – any node can commute with any other node.

ICASSP2020 Barcelona

[6] M. Milgram, "Irreducible graphs," Journal Of Combinatorial Theory (B), vol. 12, pp. 6–31, Feb. 1972..

Graph metric learning (cont'd)

Problem formulation.

- Find $\mathbf{M} \in \mathcal{S}$.

Mahalanobis distance
Type equation here.

$$\min_{\mathbf{M} \in \mathcal{S}} Q\left(\{\delta_{i,j}(\mathbf{M})\}\right), \quad \text{s.t. } \operatorname{tr}(\mathbf{M}) \le C \longrightarrow C > 0$$

avoid the trace of $\mathbf{M}$ being infinity.

convex and differentiable function

- Initialize $\mathbf{M}^0$.

$$\begin{cases} m^0_{i,i} := C/K \\ m^0_{i,j|j=i\pm 1} := -\epsilon \\ m^0_{i,j|j\neq i\pm 1} := 0 \end{cases}$$

$$\begin{bmatrix} 1 & -0.01 & 0 & 0 \\ -0.01 & 1 & -0.01 & 0 \\ 0 & -0.01 & 1 & -0.01 \\ 0 & 0 & -0.01 & 1 \end{bmatrix}$$

$\mathbf{M}^0$

Graph metric learning (cont'd)

- Optimization of diagonal terms.

$$\min_{\{m_{i,i}\}} \; Q(\mathbf{M})$$

$$\text{s.t.} \quad \mathbf{M} \succ 0; \quad \sum_i m_{i,i} \le C; \quad m_{i,i} > 0, \; \forall i.$$

GCT

$$m_{i,i} \geq \sum_{j \,|\, j \neq i} |m_{i,j}| + \rho, \qquad \forall i \in \{1, \ldots, K\}$$

$$\rho > 0$$

Linear constraints

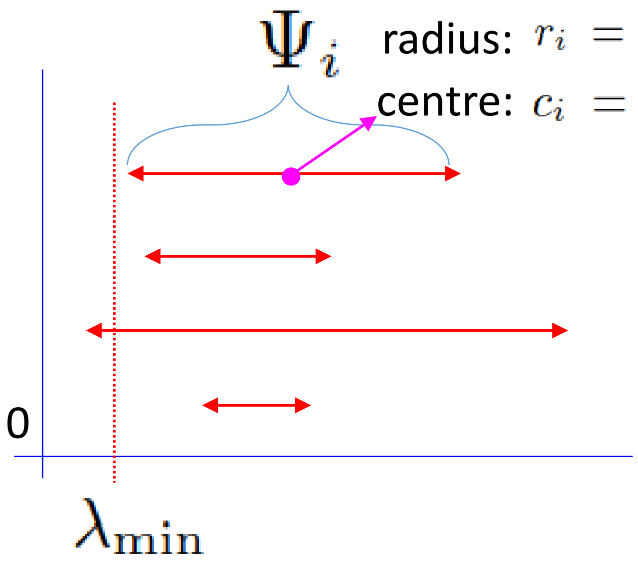
Search space is much ***smaller*** than $\mathbf{M} \succ 0$!!

Gershgorin Circle Theorem (GCT) [7]

Each eigenvalue λ of $\mathbf{M}$ resides in *at least* one **Gershgorin disc** Ψ_i

radius: $r_i = \sum_{j \,|\, j \neq i} |m_{i,j}|$

centre: $c_i = m_{i,i}$



ICASSP2020 Barcelona

[7] R. S. Varga, Gershgorin and his circles. Springer, 2004.

Graph metric learning (cont'd)

- Optimization of diagonal terms.

➢ Examine Gershgorin discs of $\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1}$, $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_K)$.

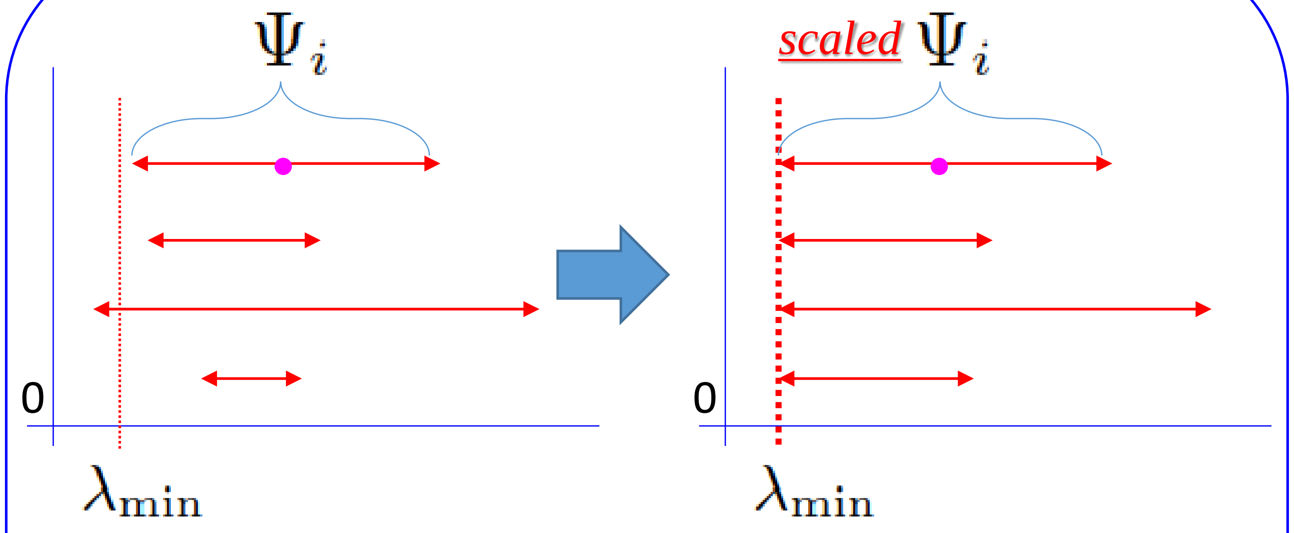
$s_k = 1/v_k$

First eigenvector $\mathbf{v}$ of $\mathbf{M}$ [8].

$\mathbf{B}$ has the same eigenvalues as $\mathbf{M}$.

same smallest Gershgorin disc's ***left-end***.

$$m_{i,i} \geq s_i \sum_{j \mid j \neq i} \frac{|m_{i,j}|}{s_j} + \rho, \quad \forall i \in \{1, \ldots, K\}$$



Theorem: these exist scalars $s_1, \ldots, s_K$ such that all Gershgorin disc left-ends are aligned at the same value $\lambda_{\min}$.

[8] C. Yang, G. Cheung, and W. Hu, "Graph Metric Learning via Gershgorin Disc Alignment," arXiv, 2020.

Graph metric learning (cont'd)

- Optimization of diagonal terms.

$$\min_{\{m_{i,i}\}} \quad Q(\mathbf{M})$$
$$\text{s.t.} \quad \mathbf{M} \succ 0; \quad \sum_i m_{i,i} \leq C; \quad m_{i,i} > 0, \forall i.$$

Gershgorin-based Reformulation

$$\min_{\{m_{i,i}\}} Q\left(\mathbf{M}\right)$$
$$\text{s.t.} \;\; m_{i,i} \geq s_i \sum_{j \,|\, j \neq i} \frac{|m_{i,j}|}{s_j} + \rho, \forall i; \quad \sum_i m_{i,i} \leq C$$

$$\min_{\{m_{i,i}\}} \mathrm{vec}(\{m_{i,i}\})^\top \, \nabla Q(\mathbf{M}^t)$$
$$\text{s.t.} \;\; m_{i,i} \geq s_i \sum_{j \,|\, j \neq i} \frac{|m^t_{i,j}|}{s_j} + \rho, \;\; \forall i; \quad \sum_i m_{i,i} \leq C.$$

- Frank-Wolfe algorithm by computing $\nabla Q(\mathbf{M}^t)$ w.r.t $\{m_{i,i}\}$.

Graph metric learning (cont'd)

- Optimization of off-diagonal entries.
- Block coordinate descent.
- Ensure irreducibility (the graph remains connected).

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & \mathbf{M}_{1,2} \\ \mathbf{M}_{2,1} & \mathbf{M}_{2,2} \end{bmatrix} \Rightarrow \min_{\mathbf{M}_{2,1}} Q(\mathbf{M})$$

$$\text{s.t. } m_{i,i} \geq s_i \sum_{j \mid j \neq i} \frac{|m_{i,j}|}{s_j} + \rho, \ \forall i$$

$$m_{\zeta,1} \leq -\epsilon; \quad \mathbf{M}_{2,1} \leq \mathbf{0}$$

At least one off-diagonal term in column 1 has magnitude at least $\epsilon > 0$.

positive edge weights

The index of previously optimized $\mathbf{M}^t_{2,1}$ with the largest magnitude

- Frank-Wolfe algorithm by computing $\nabla Q(\mathbf{M}^t)$ w.r.t $\mathbf{M}^t_{2,1}$.

Results

- Objective $Q(\mathbf{M})$: Graph Laplacian regularizer (GLR) [9].

graph signal

Node pairs

$$Q(\mathbf{M}) = \mathbf{z}^\top \mathbf{L}(\mathbf{M})\mathbf{z} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(z_i - z_j)^2$$

Edge weight

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} \exp\left\{-(\mathbf{f}_i - \mathbf{f}_j)^\top \mathbf{M}(\mathbf{f}_i - \mathbf{f}_j)\right\}(z_i - z_j)^2.$$

- Small GLR:
  - signal $\mathbf{z}$ at connected similar pairs (z_i, z_j) has a large $w_{i,j}$.
  - $\mathbf{z}$ is ***smooth*** w.r.t the variation operator $\mathbf{L}(\mathbf{M})$.

[9] J. Pang and G. Cheung, “Graph Laplacian regularization for image denoising: Analysis in the continuous domain,” IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 1770–1785, April 2017.

Results (cont'd)

- Evaluate performance in ***classification*** tasks.

- Datasets:
  1) iris (3 classes,4 features and 150 samples).
  2) wine (3 classes, 13 features and 178 samples).
  3) seeds (3 classes, 7 features and 210 samples).

- Competing schemes:
  1) learning the *diagonal terms only*: ICML'03 [10], APSIPA'16 [11], APSIPA'18 [12].
  2) learning the **full** metric matrix: ICML'16 [13], TSP'20 [14].

[10] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in ICML, Aug. 2003, pp. 912–919.
[11] Y. Mao, G. Cheung, C.-W. Lin, and Y. Ji, "Joint learning of similarity graph and image classifier from partial labels," in APSIPA, Dec. 2016, pp. 1–4.
[12] C. Yang, G. Cheung, and V. Stankovic, "Alternating binary classifier and graph learning from partial labels," in APSIPA, Nov. 2018, pp. 1137–1140.
[13] P. Zadeh, R. Hosseini, and S. Sra, "Geometric mean metric learning," in ICML, June 2016, pp. 2464–2471.
[14] W. Hu, X. Gao, G. Cheung, and Z. Guo, "Feature graph learning for 3d point cloud denoising," to apper, IEEE TSP, 2020.

Results (cont'd)

Methods		iris		wine		seeds	
		kNN	Graph-based	kNN	Graph-based	kNN	Graph-based
Diagonal-only	ICML'03 [10]	4.61	4.41	3.84	4.88	7.30	7.20
	APSIPA'16 [11]	4.97	4.57	4.61	5.18	7.15	6.93
	APSIPA'18 [12]	5.45	5.49	4.35	4.96	7.78	7.40
Full matrix	ICML'16 [13]	6.12	10.40	**3.58**	4.37	**6.92**	6.63
	TSP'20 [14]	**4.35**	4.80	4.12	4.36	7.77	7.47
	Prop.	**4.35**	**4.12**	4.27	**4.19**	7.10	**6.61**

[10] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in ICML, Aug. 2003, pp. 912–919.
[11] Y. Mao, G. Cheung, C.-W. Lin, and Y. Ji, "Joint learning of similarity graph and image classifier from partial labels," in APSIPA, Dec. 2016, pp. 1–4.
[12] C. Yang, G. Cheung, and V. Stankovic, "Alternating binary classifier and graph learning from partial labels," in APSIPA, Nov. 2018, pp. 1137–1140.
[13] P. Zadeh, R. Hosseini, and S. Sra, "Geometric mean metric learning," in ICML, June 2016, pp. 2464–2471.
[14] W. Hu, X. Gao, G. Cheung, and Z. Guo, "Feature graph learning for 3d point cloud denoising," to apper, IEEE TSP, 2020.

Thank you!

genec@yorku.ca