

Generalized Maximum Margin Clustering & Unsupervised Kernel Learning

Hamed Valizadegan , *Rong Jin*

Department of Computer Science and Engineering
Michigan State University, USA

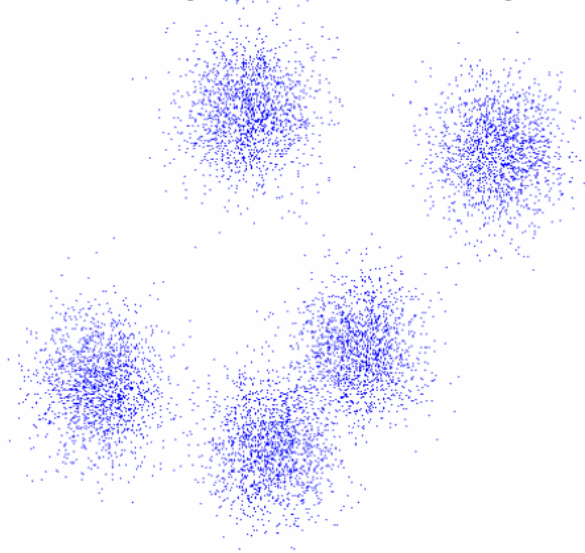


Outline

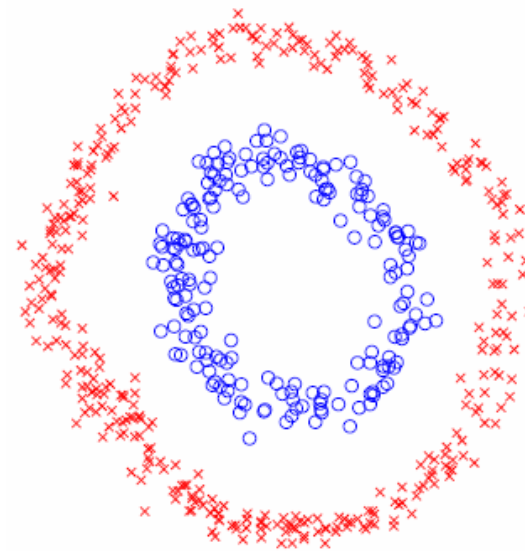
- Maximum margin data clustering
- Generalized maximum margin clustering
- Unsupervised kernel learning
- Empirical studies
- Future work

Data Clustering

- Two different criteria
 - Compactness, e.g., k-means, mixture models
 - Connectivity, e.g., spectral clustering, maximum margin clustering



Compactness



Connectivity

Maximum Margin Clustering (MMC)

- Key idea: extend SVM for unsupervised learning

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K_{i,j} \\ \text{s. t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n \end{aligned}$$

The dual problem of SVM

Training examples:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

Example weights:

$$\alpha_i, i = 1, \dots, n$$

Kernel similarity:

$$K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$$

C : weight for
classification error

Maximum Margin Clustering (MMC)

- Key idea: extend SVM for unsupervised learning

$$\begin{aligned} \max_{\alpha_i, y_i \in \{-1, +1\}} & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K_{i,j} \\ \text{s. t.} & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n \end{aligned}$$

Training examples:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

Example weights:

$$\alpha_i, i = 1, \dots, n$$

Kernel similarity:

$$K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$$

C : weight for
classification error

The dual problem of SVM

MMC: Convex Programming

- Converting into a convex programming prob.

$$\begin{aligned}
 & \min_{\mathbf{y}, \lambda, \nu, \delta} t \\
 \text{s. t.} & \begin{pmatrix} (\mathbf{y}\mathbf{y}^\top) \circ K & \mathbf{e} + \nu - \delta + \lambda\mathbf{y} \\ (\mathbf{e} + \nu - \delta + \lambda\mathbf{y})^\top & t - 2C\delta^\top \mathbf{e} \end{pmatrix} \succeq 0 \\
 & \nu \geq 0, \delta \geq 0
 \end{aligned}$$

1. Non-convex terms
2. Related to threshold b

$A \pm K$: element wise product between matrix A and B

ν, \pm vectors of size $n \times 1$

Maximum Margin Clustering

- Key idea: convert into a convex programming prob.

$$\begin{aligned} & \min_{\mathbf{y}, \nu, \delta} t \\ \text{s. t.} & \begin{pmatrix} (\mathbf{y}\mathbf{y}^\top) \circ K & \mathbf{e} + \nu - \delta \\ (\mathbf{e} + \nu - \delta)^\top & t - 2C\delta^\top \mathbf{e} \end{pmatrix} \succeq 0 \\ & \nu \geq 0, \delta \geq 0 \end{aligned}$$

$$\mathbf{y}\mathbf{y}^\top \rightarrow M \in \mathbb{R}^{n \times n}$$

$$(1) M \succeq 0,$$

$$(2) M_{i,i} = 1, i = 1, \dots, n,$$

$$(3) \text{rank}(M) = 1$$

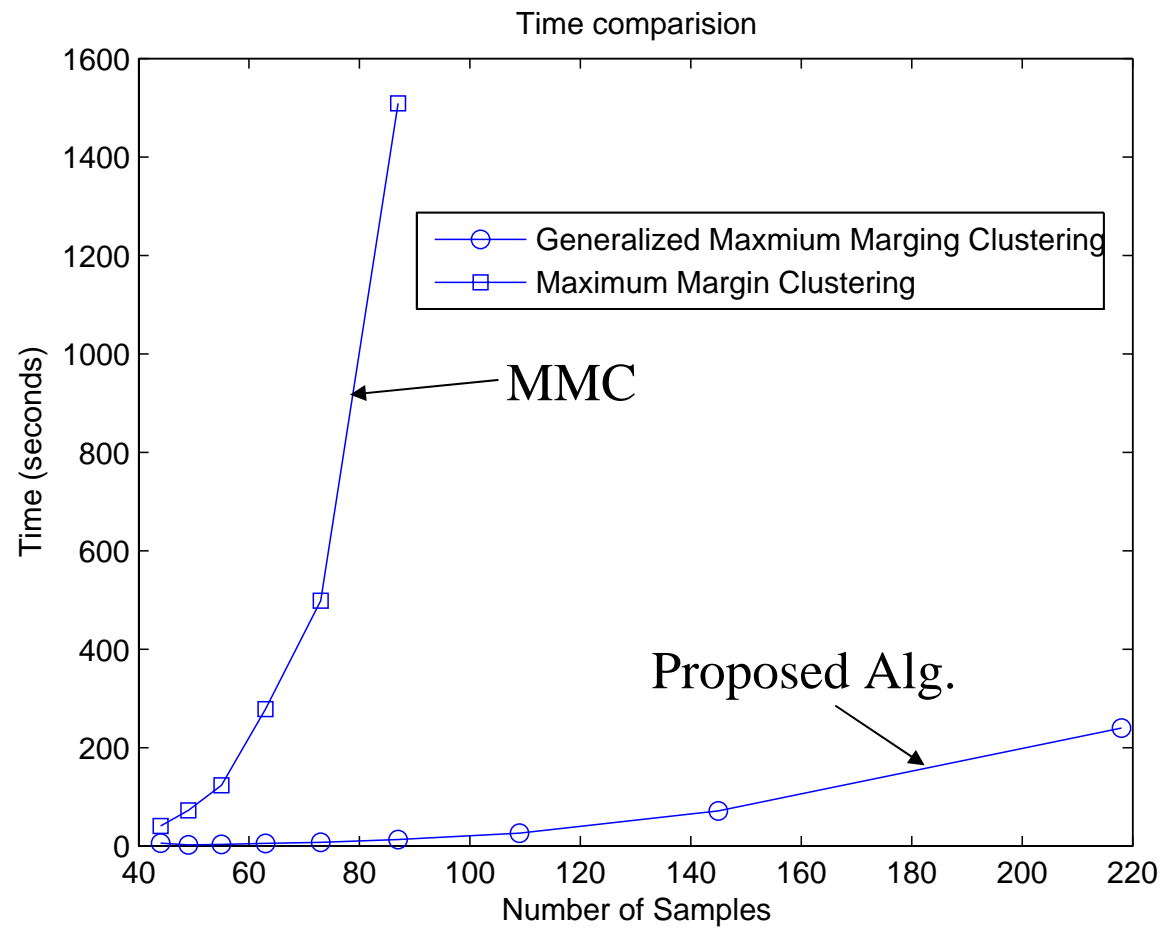
MMC: Final Formulism

$$\begin{aligned} \min_{M, \nu, \delta} \quad & t \\ \text{s. t.} \quad & \begin{pmatrix} M \circ K & \mathbf{e} + \nu - \delta \\ (\mathbf{e} + \nu - \delta)^\top & t - 2C\delta^\top \mathbf{e} \end{pmatrix} \succeq 0 \\ & \nu \geq 0, \delta \geq 0, M \succeq 0 \\ & M_{i,i} = 1, i = 1, 2, \dots, n \end{aligned}$$

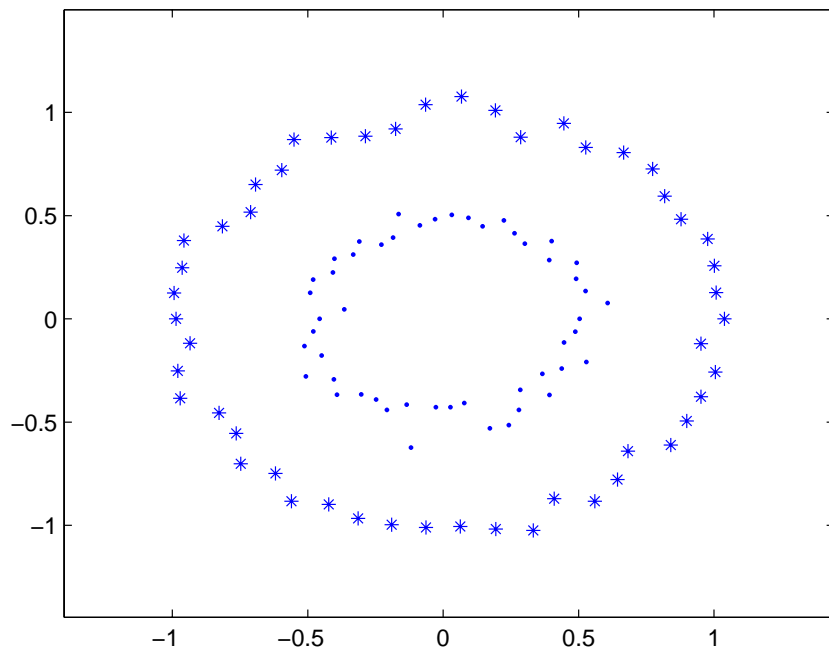
□ Disadvantages

- High computational cost: SDP, M is a $n \times n$ matrix
- Assume clustering boundaries pass through the origins.
- Sensitive to the choice of kernel functions

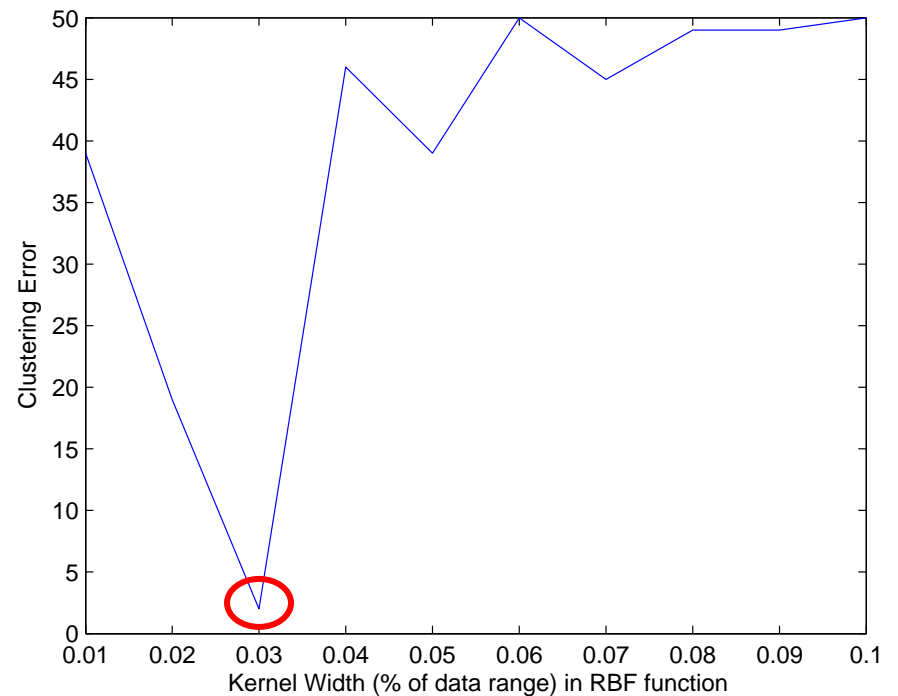
MMC: Scalability Issue



MMC: Sensitivity to Kernel Function



Dataset



MMC clustering using RBF kernel

$$K_{i,j} = \exp \left(- \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2} \right)$$

Generalized Maximum Margin Clustering

□ Simple Case: Hard Margin

$$\min_{\nu, \mathbf{y}, \lambda} \frac{1}{2} ((\mathbf{e} + \nu + \lambda \mathbf{y}) \circ \mathbf{y})^T K^{-1} ((\mathbf{e} + \nu + \lambda \mathbf{y}) \circ \mathbf{y})$$

s. t. $\nu \geq 0, \mathbf{y} \in \{+1, -1\}^n$

$$\mathbf{z} = (\mathbf{e} + \nu) \circ \mathbf{y}$$

Generalized Maximum Margin Clustering

□ Simple Case: Hard Margin

$$\min_{\nu, \mathbf{y}, \lambda} \frac{1}{2} ((\mathbf{e} + \nu + \lambda \mathbf{y}) \circ \mathbf{y})^T K^{-1} ((\mathbf{e} + \nu + \lambda \mathbf{y}) \circ \mathbf{y})$$

s. t. $\nu \geq 0, \mathbf{y} \in \{+1, -1\}^n$

↓ $\mathbf{z} = (\mathbf{e} + \nu) \circ \mathbf{y}$

$$\min_{\mathbf{z}, \lambda} \frac{1}{2} (\mathbf{z} + \lambda \mathbf{e})^T K^{-1} (\mathbf{z} + \lambda \mathbf{e})$$

s. t. $z_i^2 \geq 1, i = 1, 2, \dots, n$

$$z_i^2 = (1 + \nu_i)^2 y_i^2 = (1 + \nu_i)^2 \geq 1 \text{ since } \nu_i \geq 0$$

GMMC: Translation Invariance

$$f(\mathbf{z}, \lambda) = \min_{\mathbf{z}, \lambda} \frac{1}{2} (\mathbf{z} + \lambda \mathbf{e})^T K^{-1} (\mathbf{z} + \lambda \mathbf{e})$$
$$\text{s. t. } z_i^2 \geq 1, i = 1, 2, \dots, n$$

$$\begin{aligned} \mathbf{z}' &= \mathbf{z} + \epsilon \mathbf{e} \\ \lambda' &= \lambda - \epsilon. \end{aligned} \quad \longrightarrow \quad f(\mathbf{z}, \lambda) = f(\mathbf{z}', \lambda')$$

- Related to the non-unique solution for threshold b in SVM

GMMC: Translation Invariance

□ Remove translation invariance $\mathbf{z}^\top \mathbf{e} \approx 0$

$$\min_{\mathbf{z}, \lambda} \frac{1}{2} (\mathbf{z} + \lambda \mathbf{e})^\top K^{-1} (\mathbf{z} + \lambda \mathbf{e}) + C_e (\mathbf{z}^\top \mathbf{e})^2$$

$$\text{s. t. } z_i^2 \geq 1, i = 1, 2, \dots, n$$

$$\begin{array}{l} \Downarrow \\ \mathbf{w} = (\mathbf{z}; \lambda) \\ P = (I_n, \mathbf{e}) \end{array}$$

$$\min_{\mathbf{w} \in \mathbb{R}^{n+1}} \mathbf{w}^\top P^\top K^{-1} P \mathbf{w} + C_e (\mathbf{e}_0^\top \mathbf{w})^2$$

$$\text{s. t. } w_i^2 \geq 1, i = 1, 2, \dots, n$$

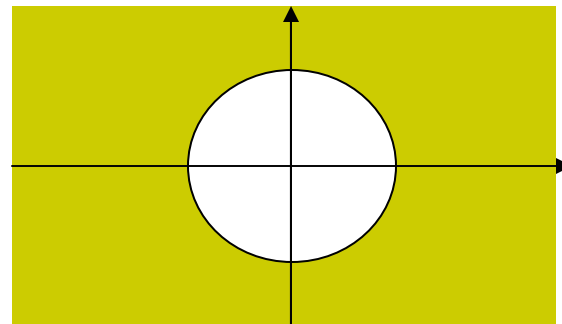
GMMC: Non-Convex Constraint

$$\min_{\mathbf{w} \in \mathbb{R}^{n+1}} \quad \mathbf{w}^T P^T K^{-1} P \mathbf{w} + C_e (\mathbf{e}_0^T \mathbf{w})^2$$

$$\text{s. t.} \quad \underline{w_i^2} \geq 1, i = 1, 2, \dots, n$$

Non-convex
Constraint

$$|\mathbf{w}|_2^2 \geq 1$$



GMMC: Dual Approximation

- Approximate the original problem by its dual

$$\max_{\gamma \in \mathbb{R}^n} \sum_{i=1}^n \gamma_i$$

$$\begin{aligned} \mathbf{e}_0 &= (\mathbf{0}_n; 1) \\ [I_{n+1}^k]_{i,j} &= \begin{cases} 1 & i = j = k \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

$$\text{s. t. } P^T K^{-1} P + C_e \mathbf{e}_0 \mathbf{e}_0^T - \sum_{i=1}^n \gamma_i I_{n+1}^i \succeq 0$$

$$\gamma_i \geq 0, \quad i = 1, 2, \dots, n$$

- The number of variables $\rightarrow n$
- A standard SemiDefinite Programming (SDP) problem

Karush-Kuhn-Tucker (KKT) Conditions

$$\left(P^T K^{-1} P + C_e \mathbf{e}_0 \mathbf{e}_0^T - \sum_{i=1}^n \gamma_i I_{n+1}^i \right) \mathbf{w} = \mathbf{0}_{n+1}$$

- \mathbf{w} is the zero eigenvector of matrix

$$\left(P^T K^{-1} P + C_e \mathbf{e}_0 \mathbf{e}_0^T - \sum_{i=1}^n \gamma_i I_{n+1}^i \right)$$



GMMC: Procedure

1. Compute $\gamma_1, \gamma_2, \dots, \gamma_n$ by solving the dual problem

GMMC: Procedure

1. Compute $\gamma_1, \gamma_2, \dots, \gamma_n$ by solving the dual problem
2. Compute the zero eigenvector \mathbf{w} of matrix

$$\left(P^T K^{-1} P + C_e \mathbf{e}_0 \mathbf{e}_0^T - \sum_{i=1}^n \gamma_i I_{n+1}^i \right)$$

3. Extract the first n elements of \mathbf{w} , and the cluster membership is determined by the sign of each element

Understand the Approximation

- Compute the dual of the dual

$$\min_M \quad \text{tr}(P^\top K^{-1} P M) + C_e \mathbf{e}_0^\top M \mathbf{e}_0$$

$$\text{s. t.} \quad M \succeq 0, \quad M_{i,i} \geq 1, \quad i = 1, 2, \dots, n$$



$$\mathbf{w} \mathbf{w}^\top \rightarrow M$$

$$(1) M \succeq 0, (2) M_{i,i} \geq 1, (3) \text{rank}(M) = 1$$

$$\min_{\mathbf{w}} \quad \mathbf{w}^\top P^\top K^{-1} P \mathbf{w} + C_e (\mathbf{e}_0^\top \mathbf{w})^2$$

$$\text{s. t.} \quad w_i^2 \geq 1, \quad i = 1, 2, \dots, n$$

Relation to Spectral Clustering

- Set $\gamma_i = \gamma, C_e \geq 1$

$$\max_{\gamma \geq 0} \gamma$$

$$\text{s. t. } K^{-1} \succeq \gamma I_n$$

- γ is the minimum eigenvalue of kernel matrix K
- Solution \mathbf{w} is related to the minimum eigenvector of $K \rightarrow$ Spectral Clustering

GMMC: Soft Margin

$$\max_{\gamma \in \mathbb{R}^n} \sum_{i=1}^n \gamma_i$$

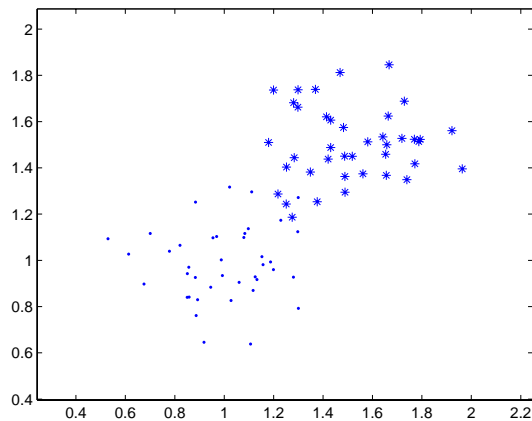
$$\text{s. t. } P^\top K^{-1} P + C_e \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^n \gamma_i I_{n+1}^i \succeq 0$$

$$0 \leq \gamma_i \leq C_\delta, \quad i = 1, 2, \dots, n$$

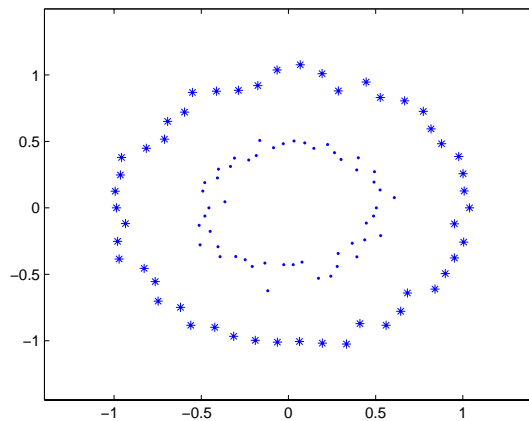
C_δ : upper bound for the assigned weight γ_i

Empirical Studies: Datasets

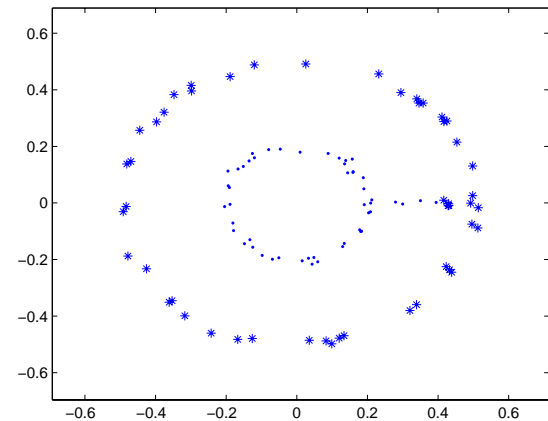
- Three synthesized datasets
 - Overlapped Gaussian, two circles, and two connected circles
- Four UCI datasets
 - Vote, digits, ionosphere, and breast



Overlapped Gaussian

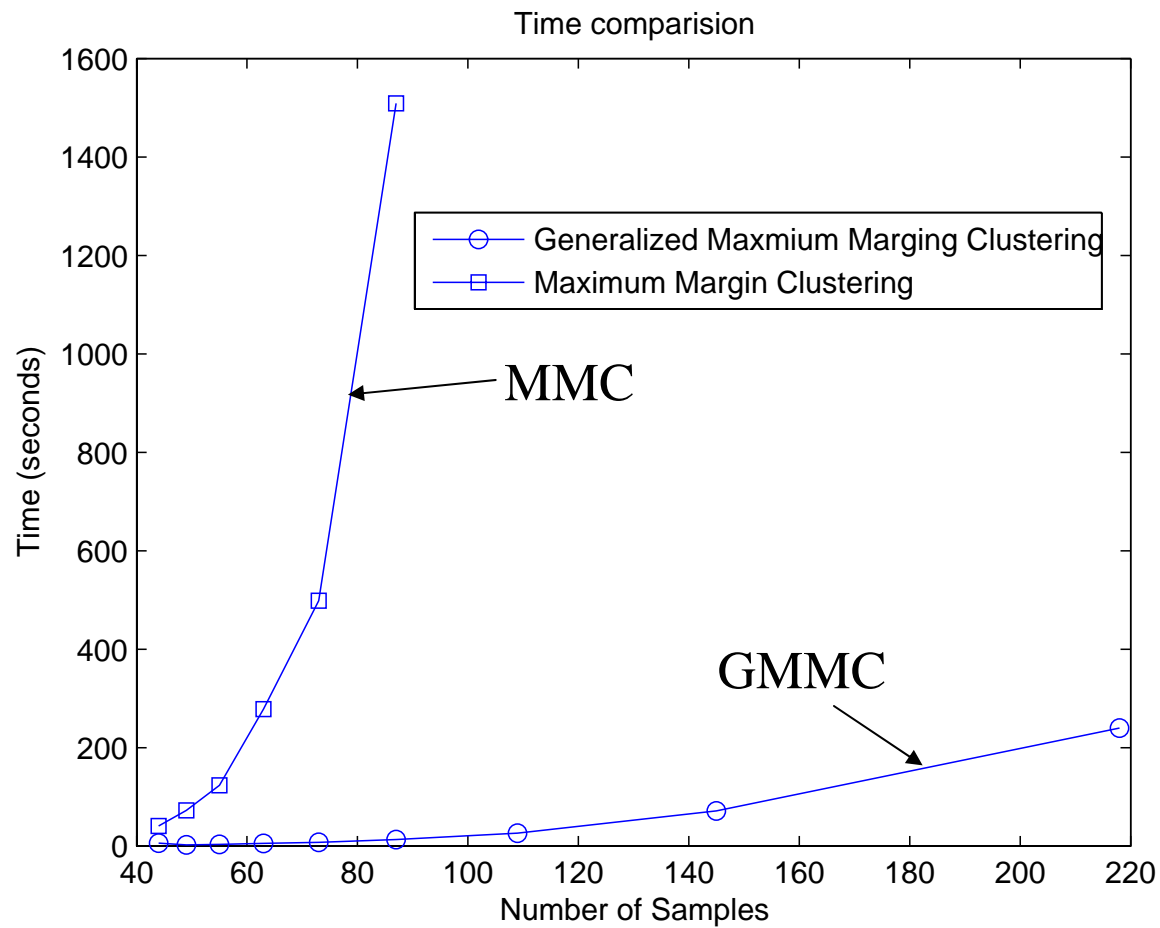


Two circles



Two connected circles

GMMC: Scalability



Empirical Studies: Results

	NC	MMC	GMMC
Two Circles	2	0	0
Two Jointed Circles	7	6.25	0
Two Gaussian	1.25	2.5	1.25
Vote	25	-	9.6
Digits 3-8	35	-	5.6
Digits 1-7	45	-	2.2
Digits 2-7	34	-	.5
Digits 8-9	48	-	16
Ionosphere	25	-	23.5
Breast	36.5	-	36.1

NC: Normalized Cut
MMC: Maximum Margin Clustering
GMMC: Generalized Maximum Margin Clustering

- Using RBF kernel with the optimal kernel width



Unsupervised Kernel Learning

- Caveat: the clustering results heavily rely on the appropriate kernel function
- How to learn the right kernel matrix without any supervision information?

Unsupervised Kernel Learning

- Given m kernel matrices $\{K_i\}_{i=1}^m$, find the right combination $K = \sum_{i=1}^m \beta_i K_i$

$$\max_{\gamma, \beta} \sum_{i=1}^n \gamma_i$$

$$\text{s. t. } P^\top \mathbf{K}^{-1} P + C_e \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^n \gamma_i I_{n+1}^i \succeq 0$$

$$0 \leq \gamma_i \leq C_\delta, \quad i = 1, 2, \dots, n$$

$$\sum_{i=1}^m \beta_i = 1, \quad \beta_i \geq 0, \quad i = 1, 2, \dots, m$$

Unsupervised Kernel Learning

- Given m kernel matrices $\{K_i\}_{i=1}^m$, find the right combination $K = \sum_{i=1}^m \beta_i K_i$

$$\begin{aligned} \max_{\gamma, \beta} \quad & \sum_{i=1}^n \gamma_i && \text{Non-convex constraint} \\ \text{s. t.} \quad & P^\top \left(\sum_{i=1}^m \beta_i K_i \right)^{-1} P + C_e \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^n \gamma_i I_{n+1}^i \succeq 0 \\ & 0 \leq \gamma_i \leq C_\delta, \quad i = 1, 2, \dots, n \\ & \sum_{i=1}^m \beta_i = 1, \quad \beta_i \geq 0, \quad i = 1, 2, \dots, m \end{aligned}$$

Unsupervised Kernel Learning

- Replace K^{-1} with combinatorial Laplacian $L(K)$

$$L(K) = D(K) - K$$

$$[D(K)]_{i,j} = \begin{cases} 0 & i \neq j \\ \sum_{k=1}^n K_{i,k} & i = j \end{cases}$$

$$K^{-1} = \left(\sum_{i=1}^m \beta_i K_i \right)^{-1} \longrightarrow L(K) = \sum_{i=1}^m \beta_i L(K_i)$$

Replacing arithmetic mean by harmonic mean

Unsupervised Kernel Learning

- Replace K^{-1} with combinatorial Laplacian $L(K)$

$$\max_{\gamma, \beta} \sum_{i=1}^n \gamma_i$$

$$\text{s. t. } P^\top \left(\sum_{i=1}^m \beta_i L_i \right) P + C_e \mathbf{e}_0 \mathbf{e}_0^\top - \sum_{i=1}^n \gamma_i I_{n+1}^i \succeq 0$$

$$0 \leq \gamma_i \leq C_\delta, \quad i = 1, 2, \dots, n$$

$$\sum_{i=1}^m \beta_i = 1, \quad \beta_i \geq 0, \quad i = 1, 2, \dots, m$$

L_i : the combinatorial Laplacian for K_i

Empirical Results

	GMMC	Self-tune (Best k)	Self-tune (Worst k)
Two Circles	0	0	50
Two Jointed Circles	0	1	45
Two Gaussian	3.75	5	7.5
Vote	11.90	11	40
Digits 3-8	5.6	5	50
Digits 1-7	3	0	47
Digits 2-7	5.6	1.5	50
Digits 8-9	12	9	48
Ionosphere	27.3	26.5	48
Breast	37	37.5	41.5

- Baseline: self-tune spectral clustering (with param k)
- Finding RBF kernel with the optimal kernel width



Conclusion

- We proposed a generalized version of maximum margin clustering
 - Reduce the computational cost
 - Remove the constraints on the clustering decision boundaries by the maximum margin clustering
 - Learn kernel matrices unsupervisedly
- Future work
 - Computational efficiency
 - Semi-supervised learning