Non-exhaustive, Overlapping Clustering via Low-Rank Semidefinite Programming

Yangyang Hou*[1], Joyce Jiyoung Whang*[2], David F. Gleich[1] and Inderjit S. Dhillon[2]
[1] Purdue University, [2] The University of Texas at Austin, * Equal Contribution
21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2015

Main Contributions

- NEO-SDP: a convex relaxation of a k-means-like objective that handles non-exhaustive, overlapping clustering problems.
- Scalable NEO-LR objective and an LRSDP algorithm to optimize a low-rank factorization of the NEO-SDP solution.
- A series of initialization and rounding strategies that accelerate the convergence of our optimization procedures.
- Evaluate LRSDP on real-world data clustering problems and find it achieves the best F_1 performance with respect to ground-truth clusters.
- For graph clustering problems, LRSDP produces the best quality communities among all clustering algorithms on real-world networks.

NEO-K-Means Objective & Iterative NEO-K-Means Algorithm

- Non-exhaustive, overlapping clustering: some data points are allowed to be outside of any cluster and clusters are allowed to overlap with each other.
- Weighted kernel NEO-K-Means objective function

$$\underset{U}{\text{minimize}} \sum_{c=1}^{k}\sum_{i=1}^{n} u_{ic} w_i \|\phi(\mathbf{x}_i) - \mathbf{m}_c\|^2, \text{ where } \mathbf{m}_c = \frac{\sum_{i=1}^{n} u_{ic} w_i \phi(\mathbf{x}_i)}{\sum_{i=1}^{n} u_{ic} w_i}$$
$$\text{subject to } \mathit{trace}(U^T U) = (1+\alpha)n,\ \sum_{i=1}^{n} \mathbb{I}\{(U\mathbf{1})_i = 0\} \le \beta n.$$

  - α and β control the degree of overlap and non-exhaustiveness.
- Weighted Kernel NEO-K-Means objective is equivalent to the extended normalized cut objective for overlapping community detection.
- The iterative NEO-K-Means Algorithm
  - Fast algorithm that monotonically decreases the NEO-K-Means objective
  - Can be trapped in local optima given poor initialization

Semidefinite Programming For NEO-K-Means

- Goal: more accurate and more reliable solutions than the iterative NEO-K-Means algorithm by paying additional computational cost
- NEO-SDP: Semidefinite Programming (SDP) for NEO-K-Means
  - Convex problem ($\rightarrow$ globally optimized via a variety solvers such as CVX)
  - Problems with fewer than 100 data points
- NEO-LR: Low-rank factorization of SDP for NEO-K-Means
  - Non-convex ($\rightarrow$ locally optimized via an augmented Lagrangian method)
  - Problems with tens of thousands of data points
- Three key variables for SDP formulations: $\mathbf{f}$ (no. of clusters each data point belongs to), $\mathbf{g}$ (indicator of non-exhaustiveness), $\boldsymbol{Z}$ (co-occurrence matrix)

$$U = \begin{array}{c c} \begin{array}{cc} C_1 & C_2 \end{array} & \\ \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} & \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \end{array} \Rightarrow \quad \mathrm{f} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix} \quad g = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \quad Z = \begin{bmatrix} \frac{{w_1}^2}{w_1+w_2} & \frac{w_1 w_2}{w_1+w_2} & 0 & 0 \\ \frac{w_2 w_1}{w_1+w_2} & \frac{{w_2}^2}{w_1+w_2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{{w_2}^2}{w_2+w_3} & \frac{w_2 w_3}{w_2+w_3} & 0 \\ 0 & \frac{w_3 w_2}{w_2+w_3} & \frac{{w_3}^2}{w_2+w_3} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

NEO-SDP

$$\begin{aligned} \underset{\boldsymbol{Z},\mathbf{f},\mathbf{g}}{\text{maximize}}\quad & \mathrm{trace}(\boldsymbol{KZ}) - \mathbf{f}^T\mathbf{d} \\ \text{subject to}\quad & \mathrm{trace}(\boldsymbol{W}^{-1}\boldsymbol{Z}) = \mathrm{k}, \\ & \boldsymbol{Z}\mathbf{e} = \boldsymbol{W}\mathbf{f}, \\ & \mathbf{e}^T\mathbf{f} = (1+\alpha)\mathrm{n}, \\ & \mathbf{e}^T\mathbf{g} \ge (1-\beta)\mathrm{n}, \\ & \mathbf{f} \ge \mathbf{g}, \\ & Z_{ij} \ge 0, \\ & \boldsymbol{Z} \succeq 0, \boldsymbol{Z} = \boldsymbol{Z}^T \\ & 0 \le \mathbf{f} \le \mathrm{k}\mathbf{e}, 0 \le \mathbf{g} \le 1 \end{aligned}$$

NEO-LR

$$\begin{aligned} \underset{\boldsymbol{Y},\mathbf{f},\mathbf{g},\mathbf{s},r}{\text{minimize}}\quad & \mathbf{f}^T\mathbf{d} - \mathrm{trace}(\boldsymbol{Y}^T\boldsymbol{K}\boldsymbol{Y}) \\ \text{subject to}\quad & \mathrm{k} = \mathrm{trace}(\boldsymbol{Y}^T\boldsymbol{W}^{-1}\boldsymbol{Y}) \\ & 0 = \boldsymbol{Y}\boldsymbol{Y}^T\mathbf{e} - \boldsymbol{W}\mathbf{f} \\ & 0 = \mathbf{e}^T\mathbf{f} - (1+\alpha)\mathrm{n} \\ & 0 = \mathbf{e}^T\mathbf{g} - (1-\beta)\mathrm{n} - r \\ & 0 = \mathbf{f} - \mathbf{g} - \mathbf{s} \\ & \mathrm{Y}_{ij} \ge 0, \\ & \mathbf{s} \ge 0, r \ge 0 \\ & 0 \le \mathbf{f} \le \mathrm{k}\mathbf{e}, 0 \le \mathbf{g} \le 1 \end{aligned}$$

- LRSDP: Solving the NEO-LR via an augmented Lagrangian method
  - Minimizing an augmented Lagrangian of the problem that includes a current estimate of the Lagrange multipliers for the constraints as well as a penalty term that drives the solution towards the feasible set.

Rounding Procedure & Practical Improvements

- Rounding procedure: getting a discrete solution from $\mathbf{f}$, $\mathbf{g}$, $\boldsymbol{Y}$
- Refinement: use LRSDP solution as the initial cluster assignment for the iterative NEO-K-Means algorithm.
- Sampling: run LRSDP on a 10% sample of the data points.
- Two-level hierarchical clustering

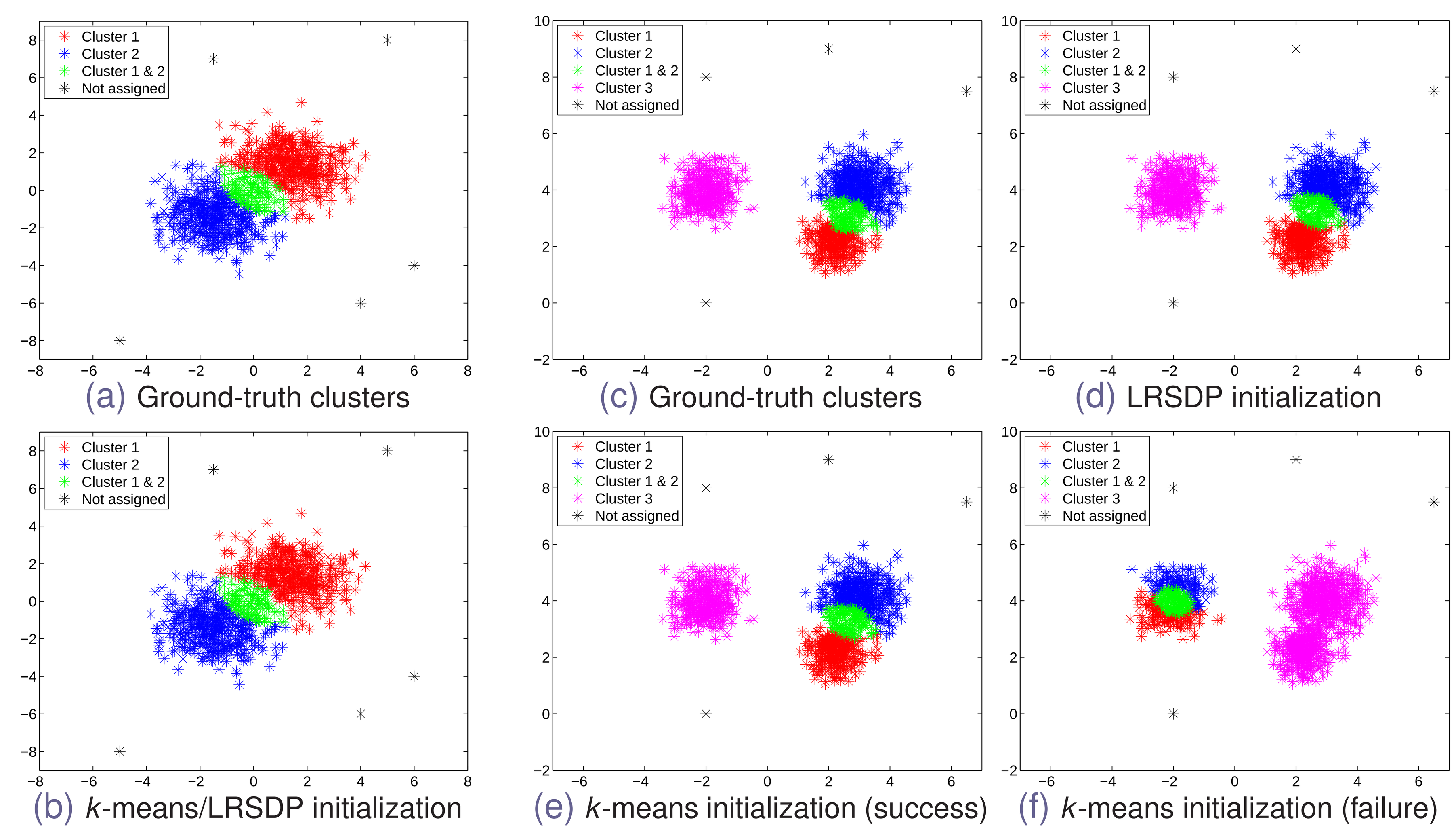
NEO-SDP via CVX vs. NEO-LR via LRSDP

- Comparison of objective values and run time
  - LRSDP is much faster than CVX, and the objective values from CVX and LRSDP are identical – they are different in light of the solution tolerances.

		Objective value		Run time (secs.)	
		SDP	LRSDP	SDP	LRSDP
dolphins	k=2, α=0.2, β=0	-1.968893	-1.968329	107.03	2.55
	k=2, α=0.2, β=0.05	-1.969080	-1.968128	56.99	2.96
	k=3, α=0.3, β=0	-2.913601	-2.915384	160.57	5.39
les miserables	k=2, α=0.2, β=0	-1.937268	-1.935365	453.96	7.10
	k=2, α=0.3, β=0	-1.949212	-1.945632	447.20	10.24
	k=3, α=0.2, β=0.05	-2.845720	-2.845070	261.64	13.53

Motivating Example: Robust LRSDP Algorithm

- NEO-K-Means algorithm with two different initializations on two datasets
  - (a), (b): On a simple dataset, NEO-K-Means can easily recover the ground-truth clusters with k-means initialization or LRSDP initialization.
  - (c)–(f): LRSDP initialization allows the NEO-K-Means algorithm to consistently produce a reasonable clustering structure whereas k-means initialization sometimes (4 times out of 10 trials) leads to a failure.



(a) Ground-truth clusters (c) Ground-truth clusters (d) LRSDP initialization
(b) k-means/LRSDP initialization (e) k-means initialization (success) (f) k-means initialization (failure)

Experimental Results on Data Clustering

- F_1 scores on real-world vector datasets
  - NEO-K-Means-based methods outperform other methods.
  - LRSDP methods improve the quality of clustering.

		moc	*esp*	*isp*	*okm*	kmeans+neo	lrsdp+neo	slrsdp+neo
yeast	worst	-	0.274	0.232	0.311	0.356	**0.390**	0.369
	best	-	0.289	0.256	0.323	0.366	**0.391**	**0.391**
	avg.	-	0.284	0.248	0.317	0.360	**0.391**	0.382
music	worst	0.530	0.514	0.506	0.524	0.526	0.537	**0.541**
	best	0.544	0.539	0.539	0.531	0.551	**0.552**	**0.552**
	avg.	0.538	0.526	0.517	0.527	0.543	0.545	**0.547**
scene	worst	0.466	0.569	0.586	0.571	0.597	**0.610**	0.605
	best	0.470	0.582	0.609	0.576	**0.627**	0.614	0.625
	avg.	0.467	0.575	0.598	0.573	0.610	**0.613**	**0.613**

Experimental Results on Overlapping Community Detection

- AUC of conductance-vs-graph coverage
  - LRSDP produces the best quality communities in terms of AUC scores.
  - The largest graph: AstroPh (17,903 nodes, 196,972 edges)

	Facebook1	Facebook2	HepPh	AstroPh
bigclam	0.830	0.640	0.625	0.645
demon	0.495	0.318	0.503	0.570
oslom	0.319	0.445	0.465	0.580
nise	0.297	0.293	0.102	0.153
multilevel neo	0.285	0.269	0.206	0.190
LRSDP	**0.222**	**0.148**	**0.091**	**0.137**

Y. Hou*, J. J. Whang*, D. F. Gleich, and I. S. Dhillon. Non-exhaustive, Overlapping Clustering via Low-Rank Semidefinite Programming. In KDD, 2015.