Constrained Spectral Clustering with L1 Regularization

Jaya Kawale and Daniel Boley

University of Minnesota

Supported by NSF grants IIS-0916750, IIS-1029711 and UofM Fellowship.

Constrained Spectral Clustering

- Partitioning of undirected graphs finds many applications in social networks, machine learning, ...
- Wish to find partition in which $\frac{\text{size of cut}}{\text{sizes of halves}}$ is small.
- Spectral Clustering is commonly used as a fast approximation, based on a quadratic cost fcn ((Shi & Malik, 2000) & others).
- A little prior knowledge can yield marked improvements in clusters (e.g. (Wagstaff et al., 2001; Yu & Shi, 2001; Ji & Xu, 2006)).
  - Prior knowledge in spectral clustering have been mostly limited to *must-link* constraints (Kamvar et al., 2003; Xu et al., 2005; Ji & Xu, 2006; Shi et al., 2010)
  - Previous method admitting *cannot-link* constraints used quadratic but indefinite cost function, and needed many eigenvectors (Wang & Davidson, 2010).

Goals

- Find an approximate minimal normalized cut while limiting the *number* of violations of known constraints.
- Handle both *must-link* and *cannot-link* constraints.
- Avoid forcing all constraints to be exactly satisfied, allowing some noise in the constraints.
- Design method that is also applicable to co-clustering.

  - Design penalty term for constraint violations that cannot dominate original quadratic cost fcn from graph.
  - Get inspiration from sparse least squares and convex relaxations of combinatoria problems.

Method

- Minimize a quadratic function (spectral cut) subject to a constraint-violatio
  penalty (count of violations).

- Relax the sparsity count to an L_1 penalty.
  - Quadratic function is the real relaxation of the normalized cut.
  - Sparsity penalty is applied to violations of *must-link* and *cannot-link* constraints.
  - Inspired by previous work in sparse least squares, like LASSO (Tibshirani, 1996), basis pursuit, compressed sensing, etc.

Issues

- Without constraints, get a [generalized] eigenvalue problem.

- With L_1 constraint penalty, get a non-convex optimization problem.

- Our simple solution: solve by a series of convex subproblems.

Spectral Clustering – Preliminaries

- Graph $G = \{V, E, W\} = \{\text{vertices, edges, edge affinities}\}$.
- Affinity between two clusters S_1, S_2 is

$$|\text{edges in cut}| = W(S_1, S_2) = \sum_{u \in S_1, v \in S_2} w_{uv}$$

- For binary cuts, normalized cut is size of cut relative to size of partitions:

$$NC_{node} = |V| \frac{W(S_1,S_2)}{|S_1| \cdot |S_2|} \qquad = |\text{vertices}| \frac{|\text{edges in cut}|}{|\text{vertices}_1| \cdot |\text{vertices}_2|}$$

$$NC_{edge} = \text{sum}(W) \frac{W(S_1,S_2)}{W(S_1,V) W(S_2,V)} = |\text{edges}| \frac{|\text{edges in cut}|}{|\text{edges}_1| \cdot |\text{edges}_2|}$$

- Both are measures of the form $\frac{\text{size of cut}}{\text{sizes of halves}}$.
- Differ in the measure of "sizes of halves": count of vertices or edges.
- For simplicity, this talk will focus on NC_{node}.

Matrix Equivalent

- Define
$$\begin{aligned} A &= \text{[weighted] adjacency matrix} \\ \mathbf{d} &= A \cdot \mathbf{1} = \text{vector of degrees} \\ D &= \text{Diag}(\mathbf{d}) \\ L &= D - A = \text{unnormalized Laplacian} \end{aligned}$$

- Then problem is:

$$\text{minimize } NC_{node} = \frac{\mathbf{x}^T L \mathbf{x}}{\mathbf{x}^T \mathbf{x}}, \qquad \text{s.t. } \mathbf{x} \perp \mathbf{1}$$

  subject to $\mathbf{x} \in \{\alpha, -\beta\}^n$ taking only 2 discrete values, with $\alpha, \beta > 0$.

- L_1 relaxation: allow $\mathbf{x}$ to take any real values.

- Resulting minimization problem to be solve:

$$\min_{\mathbf{x}} \ {}^1\!/\!_2 \mathbf{x}^T L \mathbf{x} \quad \text{s.t.} \quad \mathbf{1}^T \mathbf{x} = 0 \quad , \quad \mathbf{x}^T \mathbf{x} = 1.$$

- Usually solved as an eigenproblem $L\mathbf{x} = \lambda \mathbf{x}$:
  Seek Fiedler vector: eigenvector for smallest nonzero eigenvalue.

Must-link & *Cannot-link* Constraints

- Old methods added new quadratic penalty term for constraint violations.
- Like modifying the graph or quadratic graph cost.
- With large weight, penalty term might hide effect of original cost fcn.

Our Approach

- Mimic counting the number of violations.
- Encode constraints in a matrix C, so that $\|C\mathbf{x}\|_0$ is the count of constraint violations.
- C resembles an incidence matrix, with rows like:

$$\begin{array}{ll} (0,\ldots,0,-1,0,\ldots,0,+1,0,\ldots,0) & (\textit{must-link}) \\ (0,\ldots,0,+1,0,\ldots,0,+1,0,\ldots,0) & (\textit{cannot-link}). \end{array}$$

Optimization Problem with Constraints

- Incorporate *must-link* & *cannot-link* constraints into optimization problem:

$$\begin{array}{lll} \min_{\mathbf{x}} & \frac{1}{2}\mathbf{x}^T L \mathbf{x} & \\ \text{s.t.} & \mathbf{d}^T\mathbf{x} = 0 & \\ & C\mathbf{x} = 0 & \textit{(enforce all constraints)} \\ & \mathbf{x}^T I \mathbf{x} = 1. & \end{array}$$

- This could be solved as a generalized eigenvalue problem (Bie et al., 2004), but could be at high expense.
- Hard constraint may be too strict if underlying clustering does not match the labels well, or there is noise in the constraints.
- Wish to have trade-off between clustering and constraints.

Optimization Problem with Constraints

- Incorporate *must-link* & *cannot-link* constraints into optimization problem:

$$\begin{array}{lll} \min_{\mathbf{x}} & \frac{1}{2}\mathbf{x}^T L\mathbf{x} + \lambda\boxed{\|\mathbf{z}\|_1} & \\ \text{s.t.} & \mathbf{d}^T\mathbf{x} = 0 & \\ & C\mathbf{x} = \mathbf{z} & \textit{(enforce some constraints)} \\ & \mathbf{x}^T I\mathbf{x} = 1. & \end{array}$$

- Use $\boxed{\|\mathbf{z}\|_1}$ as a convex relaxation for the count $\|\mathbf{z}\|_0$.
- Soft constraint admits trade-off for clustering distortion or noise in the constraints.
- Even if λ is large, the original L term is never completely lost.

Convex Subproblem

- Previous problem is not convex.
- Solve by repeated solution of a convex subproblem with proximity penalty:

$$\begin{array}{ll} \min_{\widehat{\mathbf{x}},\widehat{\mathbf{z}}} & \frac{1}{2}\widehat{\mathbf{x}}^T L \widehat{\mathbf{x}} + \mu \|\widehat{\mathbf{x}} - \boxed{\mathbf{x}}\|_I^2 + \lambda \|\widehat{\mathbf{z}}\|_1 \\ \text{s.t.} & \mathbf{d}^T \widehat{\mathbf{x}} = 0 \\ & C\widehat{\mathbf{x}} - \widehat{\mathbf{z}} = 0 \\ & \boxed{\mathbf{x}}^T I \widehat{\mathbf{x}} = 1, \end{array}$$

- $\boxed{\mathbf{x}}$ is used as the starting point for subproblem.
- Proximity term keeps new iterate $\widehat{\mathbf{x}}$ close to starting $\mathbf{x}$, with wgt μ.
- Quadratic Constraint replaced with linear approximation.

Convex Subproblem

- Previous problem is not convex.
- Solve by repeated solution of a convex subproblem with proximity penalty:

$$\begin{array}{ll} \min_{\widehat{\mathbf{x}},\widehat{\mathbf{z}}} & \frac{1}{2}\widehat{\mathbf{x}}^T L\widehat{\mathbf{x}} + \boxed{\mu\|\widehat{\mathbf{x}} - \boxed{\mathbf{x}}\|_I^2} + \lambda\|\widehat{\mathbf{z}}\|_1 \\ \text{s.t.} & \mathbf{d}^T\widehat{\mathbf{x}} = 0 \\ & C\widehat{\mathbf{x}} - \widehat{\mathbf{z}} = 0 \\ & \boxed{\mathbf{x}}^T I\widehat{\mathbf{x}} = 1, \end{array}$$

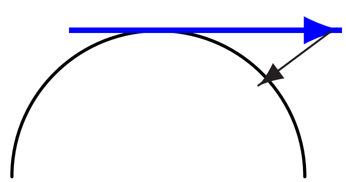
- $\boxed{\mathbf{x}}$ is used as the starting point for subproblem.
- Proximity term keeps new iterate $\widehat{\mathbf{x}}$ close to starting $\boxed{\mathbf{x}}$, with wgt μ.
- Quadratic Constraint replaced with linear approximation .

Convex Subproblem

- Previous problem is not convex.
- Solve by repeated solution of a convex subproblem with proximity penalty:

$$\begin{array}{ll} \min_{\widehat{\mathbf{x}},\widehat{\mathbf{z}}} & \frac{1}{2}\widehat{\mathbf{x}}^T L \widehat{\mathbf{x}} + \boxed{\mu \|\widehat{\mathbf{x}} - \boxed{\mathbf{x}}\|_I^2} + \lambda \|\widehat{\mathbf{z}}\|_1 \\ \text{s.t.} & \mathbf{d}^T \widehat{\mathbf{x}} = 0 \\ & C\widehat{\mathbf{x}} - \widehat{\mathbf{z}} = 0 \\ & \boxed{\boxed{\mathbf{x}}^T I \widehat{\mathbf{x}} = 1,} \end{array}$$

- $\boxed{\mathbf{x}}$ is used as the starting point for subproblem.
- Proximity term keeps new iterate $\widehat{\mathbf{x}}$ close to starting $\boxed{\mathbf{x}}$, with wgt μ.
- Quadratic Constraint replaced with linear approximation.



Overall Algorithm

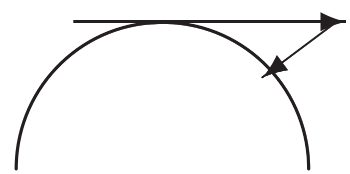
Start with Laplacian L, constraints C, scalars λ, μ, initial $\mathbf{x}^{[0]}$.

1. For $k = 0, 1, 2, \ldots$ until convergence
2. Solve convex subproblem for $\widehat{\mathbf{x}}_{\text{min}}, \widehat{\mathbf{z}}_{\text{min}}$, starting with $\mathbf{x} = \mathbf{x}^{[k]}$
3. Set $\gamma = \|\widehat{\mathbf{x}}^T_{\text{min}}\|_I$
4. Set $\mathbf{x}^{[k+1]} = \widehat{\mathbf{x}}_{\text{min}}/\gamma$
5. Set $\mathbf{z}^{[k+1]} = \widehat{\mathbf{z}}_{\text{min}}/\gamma$

(steps 3–5:) (*project back onto sphere* $\mathbf{x}^T I \mathbf{x} = 1$)

Return: $\mathbf{x}^{[\text{final}]}$: cluster indicator vector.

- Theorem: each pass through subproblem is a descent step for original problem.



Experimental Setup

- Use some simple datasets with samples in $\mathbb{R}^n$ and known labels.
- Construct pair-wise affinity matrix using $A_{ij} = \exp\left(-\frac{1}{2\sigma}\|x_i - x_j\|_2^2\right)$.
- Measure performance with cluster Purity and Normalized Mutual Informatior
    - $Purity(\widehat{\mathbf{x}}, \mathbf{y}) = \sum_k max_j \left\{ \frac{|c_k \cap l_j|}{|c_k|} \right\}$
      (fraction of most common label within each cluster (Zhao & Karypis, 2004)).
    - $NMI(\widehat{\mathbf{x}}, \mathbf{y}) = \frac{2 \cdot I(\widehat{\mathbf{x}}, \mathbf{y})}{H(\widehat{\mathbf{x}}) + H(\mathbf{y})}$
      (Normalized Mutual Information (Zhong & Ghosh, 2005)).
    - Ranges: $Purity \in [{}^1\!/\!_2, 1]$, $NMI \in [0, 1]$, with $1 =$ perfect match.
- Compare with Baseline method admitting both *Must-link* & *Cannot-link* constraints.

Baseline Method

- Found only one baseline method capable of handling *Cannot-link* constraints (Wang & Davidson, 2010).
- Use $Q = \{-1, 0, 1\}^{n \times n}$ (constraints),
- Form modified generalized eigenvalue problem using L and Q.
- Solution of eigenvalue problem is expensive.
    - $\widetilde{L} = D^{-1/2} L D^{-1/2}$, $\widetilde{Q} = D^{-1/2} Q D^{-1/2}$.
    - Leads to: $\min_{\mathbf{v}} \mathbf{v}^T \widetilde{L} \mathbf{v}$ s.t. $\mathbf{v}^T \widetilde{Q} \mathbf{v} = \alpha$, $\mathbf{v}^T \mathbf{v} = vol, \mathbf{v} \neq \sqrt{\mathbf{d}}$
    - Solve by selecting from eigenvectors of $\widetilde{L}\mathbf{v} = \lambda(\widetilde{Q} - \frac{\beta}{vol} I)\mathbf{v}$ for $\lambda > 0$.
    - β is user-supplied. Need to compute all eigenvectors (expensive).

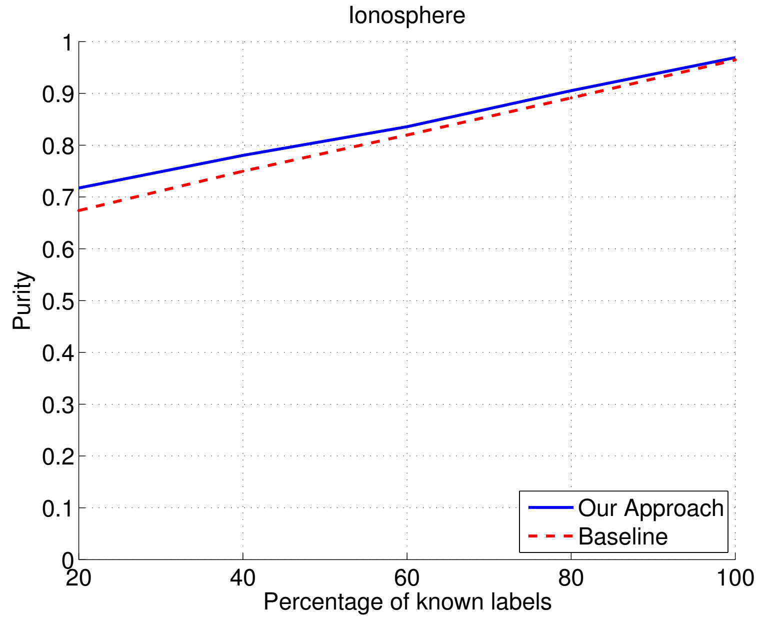
Small Experimental Datasets

Data	No of instances	No of attributes
Wine	119	13
Glass	146	9
Ionosphere	351	32
Hepatitis	155	19
WDBC	569	30
Diabetes	768	8

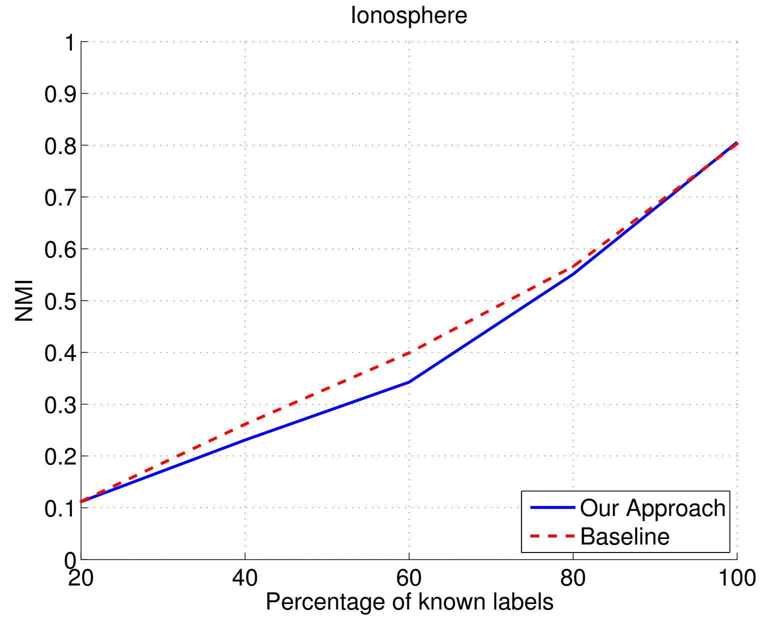
- Graph constructed using RBF kernel on pair-wise distances.

Ionosphere

Purity

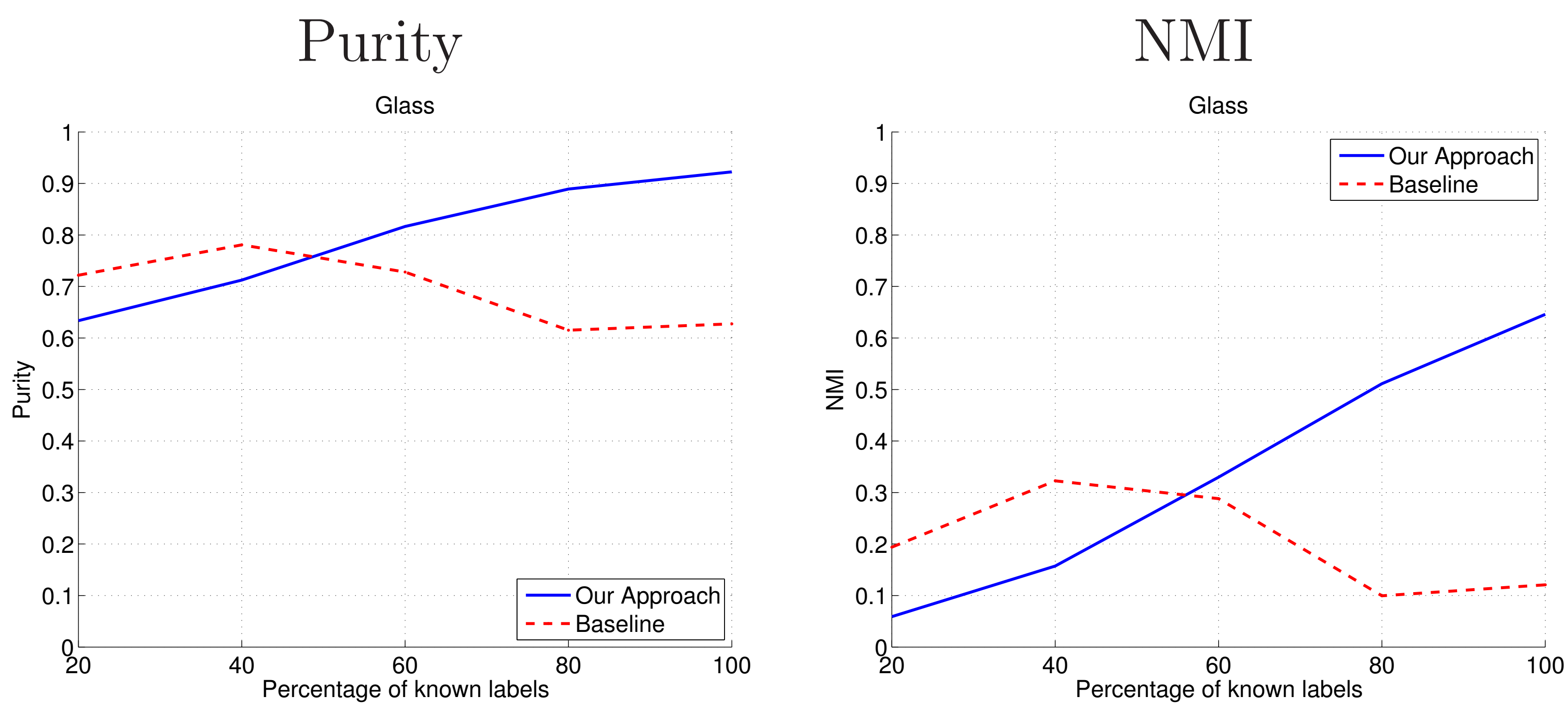


NMI



- Typical behavior on easy (well separated) datasets.
- Sometimes our method slightly better, sometimes baseline slightly better.
- Any method does well.

Glass



- When the natural clustering fails to capture the labels, our method is better able to follow the constraints when there are enough of them.

Performance – Fixed NMI

- α value needed by baseline method to achieve given NMI
- α may go almost off-scale, burying the original Laplacian.
- In all cases $\lambda \in [.1, 10]$ for our method.

Dataset	$NMI \in [.6, .7]$	$NMI \in [.7, .8]$	$NMI \in [.8, .9]$	$NMI \in [.9, 1]$
Wine	–	56.56	88447	2.1624e+05
Glass	–	–	–	–
Ionosphere	0.62	0.82	0.9294	1.0325
Hepatitis	6.19e+07	5.36e+09	6.42e+09	6.424e+09
WDBC	–	1.99e+03	6.77e+14	6.16e+23
Diabetes	217	485.50	2.12e+03	2.28e+03

Performance – Satisfy Fixed % Labels

- α value needed by baseline to match a pre-set % of given labels.
- α may go almost off-scale, burying the original Laplacian.
- In all cases $\lambda \in [.1, 10]$ for our method.

Dataset	$\%known$ 20	$\%known$ 40	$\%known$ 60	$\%known$ 80
Wine	1.85e+02	1.33e+03	8.85e+04	3.03e+05
Glass	3.27e+06	3.27e+06	3.29e+06	4.34e+34
Ionosphere	0.20	0.41	0.61	0.82
Hepatitis	6.19e+07	1.20e+05	6.36e+09	6.424e+09
WDBC	6.77e+14	1.53e+14	6.16e+23	6.16e+23
Diabetes	279	346	2.12e+03	2.28e+03

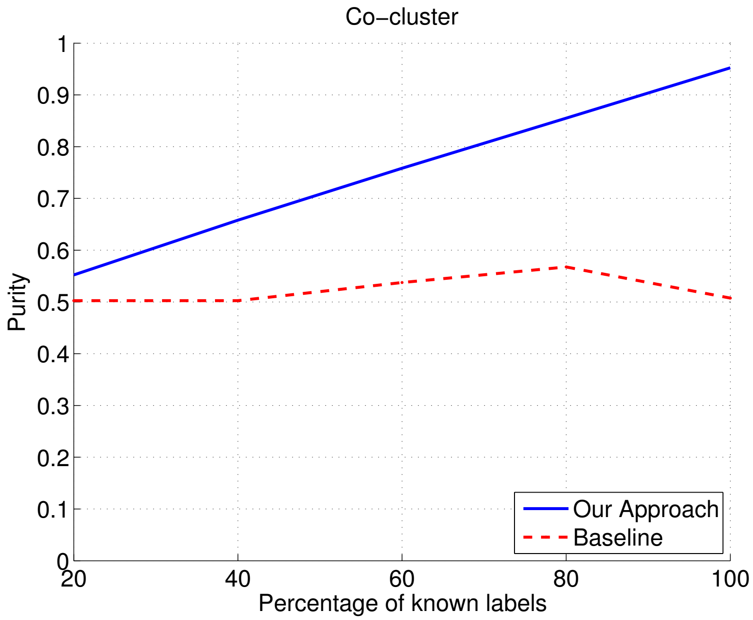
Co-Clustering Dataset

Data	No. of Docs	No. of edges
`Medline`	200	10510
`Cranfield`	200	10210
Total	400	20720
combined bipartite graph	No. of nodes	No. of edges
	3514	20720

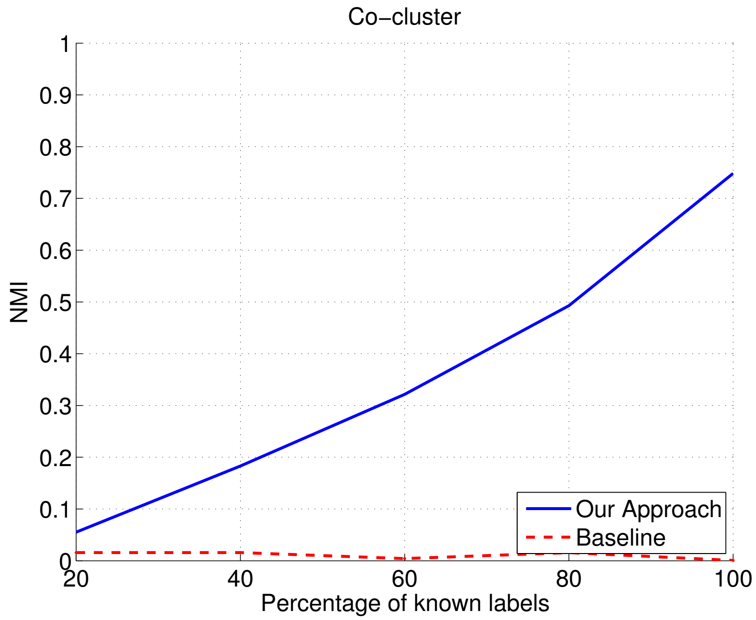
- Co-clustering can often do better than ordinary clustering when both attributes (words) and samples (docs) separate.
- Use bipartite graph connecting words to documents.
- We combined two separate datasets into a single bipartite graph.

Co-Clustering Results

Purity



NMI

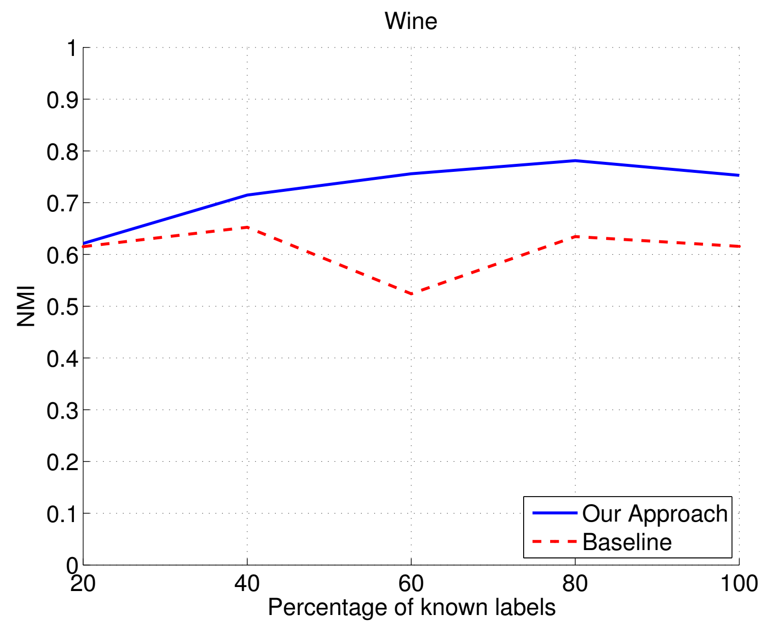


- Our method follows constraints; baseline method does not.

Noisy Labels (10%)

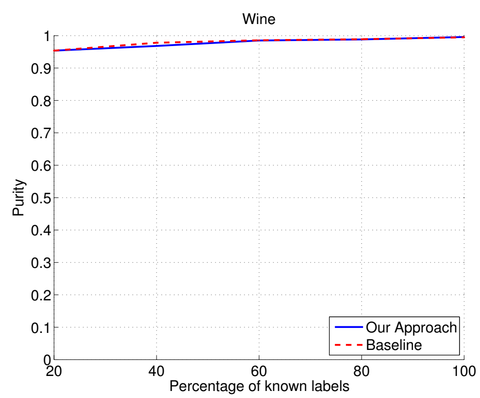
- 10% of the constraints were randomly flipped.
- Goal: simulate noise in the data.
- Our method better able to use imperfect prior knowledge compared to baseline method.

NMI

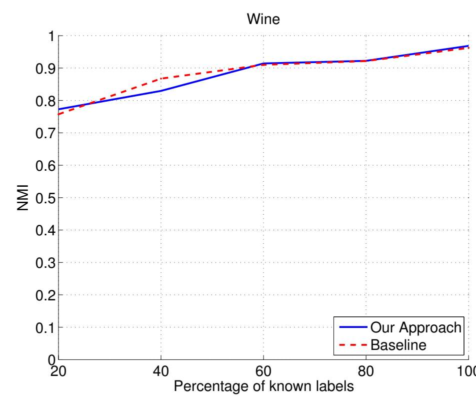


- Noise-free performance shows this data is well separated:

Purity



NMI



Slightly Larger Example

- Selected 0's & 1's (1389 samples) from USPS digit dataset.
  - Each sample is a 16 by 16 image converted to a 256-vector.
  - Form unweighted graph by connecting each sample to samples within 20% of maximum distance.
  - Added at least two edges to a sample far away.
  - Total number of edges 431516 with degrees ranging from 5 to 1079.
  - 4316 Constraints: 1% of all possible pair-wise agreements/disagreements.
- Natural spectral clustering: 118 disagreements with ground-truth labels
- After 11 iterations of convex subproblem (total time: 373 sec) disagreements reduced to 30.

Conclusions

- We have presented a way to incorporate *Must-link* & *Cannot-link* constraints into spectral [co-]clustering.
- We use an L1 penalty term on the constraints to avoid overwhelming the underlying affinity graph.
- We showed how the non-convex problem can be solved by a sequence of convex subproblems which includes a proximity penalty.
- We illustrated that this method can be robust in the presence of noise in the constraints.

Future Directions – Wishlist

- A more efficient solver for the convex subproblem, or the original non-convex problem.
- Extension to more than two clusters (perhaps by recursive binary splitting).
- Exploration of the choice of parameters, including fast tracking as λ varies.

Thank you!

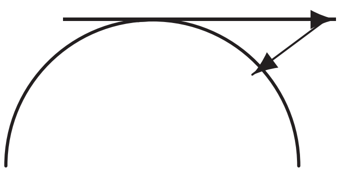
Algorithm Convergence

Theorem Each pass through steps 2–5 of Algorithm is a descent step for original non-convex optimization problem.

Proof (sketch)

1. Convex subproblem reduces original objective function.

2. Length $\gamma = \|\widehat{\mathbf{x}}^T_{\min}\|_D > 1$.



3. Scaling by $1/\gamma$ further reduces original objective function, while landing on original feasible region.

Choice of Parameters

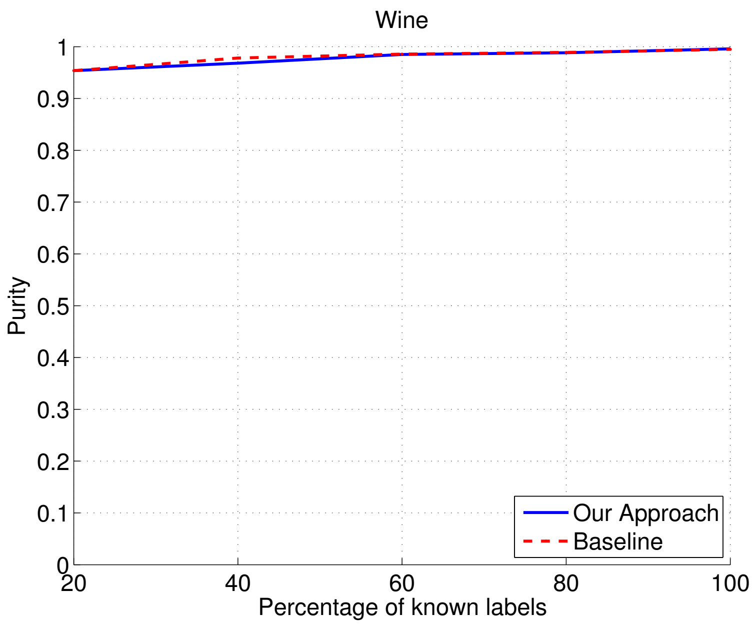
- Choice of constraint weight: $\lambda = [0.1, 10]$ worked well for all cases tried.
- proximity penalty $\mu = 1$ was a good balance between quadratic cost function and the quadratic proximity penalty.
- Subproblem converged in 6-8 iterations in most cases.

Measuring cluster quality

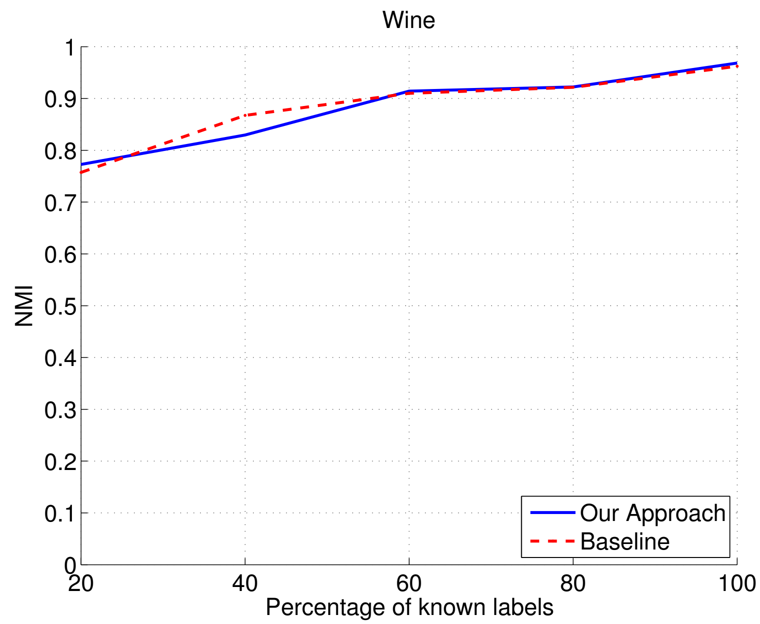
- Cluster quality measured by comparing with labels (ground truth).
- In general, it is hard to measure how well the natural affinities in the graph are aligned with a given set of labels.
- $Purity(\widehat{\mathbf{x}}, \mathbf{y}) = \sum_k max_j \left\{ \frac{|c_k \cap l_j|}{|c_k|} \right\}$ = fraction of most common label within each cluster (Zhao & Karypis, 2004).
- $NMI(\widehat{\mathbf{x}}, \mathbf{y}) = \frac{2 \cdot I(\widehat{\mathbf{x}}, \mathbf{y})}{H(\widehat{\mathbf{x}}) + H(\mathbf{y})}$ = Normalized Mutual Information (Zhong & Ghosh, 2005).
- $Purity \in [{}^1\!/_2, 1]$, $NMI \in [0, 1]$, with $1 =$ perfect match.

Wine

Purity



NMI



- Typical behavior on easy (well separated) datasets.
- Any method does well.

References

Bie, T. D., Suykens, J. A. K., & Moor, B. D. (2004). Learning from general label constraints. *SSPR/SPR* (pp. 671–679).

Ji, X., & Xu, W. (2006). Document clustering with prior knowledge. *SIGIR* (pp. 405–412).

Kamvar, S. D., Klein, D., & Manning, C. D. (2003). Spectral learning. *IJCAI* (pp. 561–566).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *PAMI*, *22*, 888–905.

Shi, X., Fan, W., & Yu, P. S. (2010). Efficient semi-supervised spectral co-clustering with constraints. *ICDM* (pp. 1043–1048).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. *ICML* (pp. 577–584).

Wang, X., & Davidson, I. (2010). Flexible constrained spectral clustering. *KDD* (pp. 563–572).

Xu, Q., desJardins, M., & Wagstaff, K. (2005). Active constrained clustering by examining spectral eigenvectors. *Discovery Science* (pp. 294–307).

Yu, S. X., & Shi, J. (2001). Grouping with bias. *NIPS* (pp. 1327–1334).

Zhao, Y., & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, *55*, 311–331.

Zhong, S., & Ghosh, J. (2005). Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, *8*, 374–384.