SoF: Soft-Cluster Matrix Factorization for Probabilistic Clustering

Han Zhao[†], Pascal Poupart[†], Yongfeng Zhang[‡] and Martin Lysy[†]
[†]{han.zhao, ppoupart, mlysy}@uwaterloo.ca, [‡]zhangyf07@gmail.com
[†]University of Waterloo, [‡]Tsinghua University

Introduction

- Probabilistic clustering without making explicit assumptions on data density distributions.
- Axiomatic approach to define 4 properties that the probability of co-clustered pairs of points should satisfy.
- Establish a connection between probabilistic clustering and constrained symmetric Nonnegative Matrix Factorization (NMF).
- Sequential minimization algorithm using penalty method to convert the constrained optimization problem into an unconstrained problem and solve it using gradient descent.

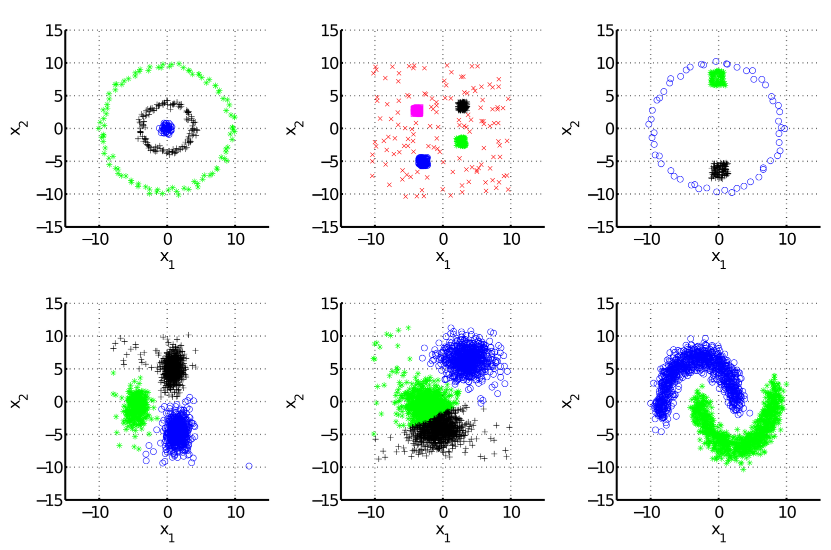

Figure 1: Synthetic experiments on 6 data sets. Different colors for different clusters found by SoF.

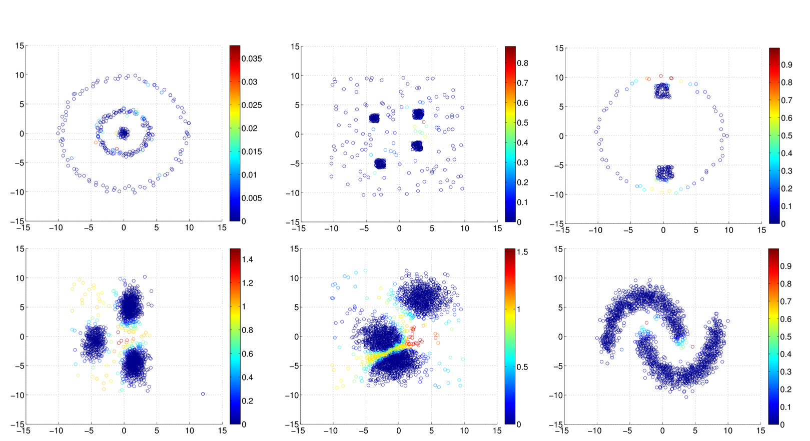

Figure 2: Entropy graph of probabilistic assignment. Brighter colors correspond to higher entropy and hence the uncertainty in clustering.

Soft-Cluster Matrix Factorization

Co-cluster Probability $p_C(\mathbf{v}_1, \mathbf{v}_2)$ induced by distance function $L : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_+$ should satisfy the following 4 properties:

- Boundary property: $\forall \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$, if $L(\mathbf{v}_1, \mathbf{v}_2) = 0$, then $p_C(\mathbf{v}_1, \mathbf{v}_2) = 1$; if $L(\mathbf{v}_1, \mathbf{v}_2) \to \infty$, then $p_C(\mathbf{v}_1, \mathbf{v}_2) \to 0$.
- Symmetry property: $\forall \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$, $p_C(\mathbf{v}_1, \mathbf{v}_2) = p_C(\mathbf{v}_2, \mathbf{v}_1)$.
- Monotonicity property: $\forall \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{R}^d$, $p_C(\mathbf{v}_1, \mathbf{v}_2) \leq p_C(\mathbf{v}_1, \mathbf{v}_3) \Leftrightarrow L(\mathbf{v}_1, \mathbf{v}_2) \geq L(\mathbf{v}_1, \mathbf{v}_3)$.
- Tail property: $\forall \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{R}^d$, $\left|\frac{\partial p_C(\mathbf{v}_1,\mathbf{v}_2)}{\partial L(\mathbf{v}_1,\mathbf{v}_2)}\right| \leq \left|\frac{\partial p_C(\mathbf{v}_1,\mathbf{v}_3)}{\partial L(\mathbf{v}_1,\mathbf{v}_3)}\right| \Leftrightarrow L(\mathbf{v}_1, \mathbf{v}_2) \geq L(\mathbf{v}_1, \mathbf{v}_3)$.

Given a distance function L, there exists a family of co-cluster probability functions $p_C(\mathbf{v}_1, \mathbf{v}_2) = e^{-cL(\mathbf{v}_1,\mathbf{v}_2)}$ which satisfy the boundary, symmetry, monotonicity and tail properties simultaneously for any constant $c > 0$.

Empirical co-cluster probability: $P_{ij} \triangleq p_C(\mathbf{v}_i, \mathbf{v}_j) = e^{-L(\mathbf{v}_i,\mathbf{v}_j)}$

True co-cluster probability: $\Pr(\mathbf{v}_i \sim \mathbf{v}_j) = \Sigma_{k=1}^K \Pr(c_i = k|\mathbf{v}_i) \times \Pr(c_j = k|\mathbf{v}_j) = \mathbf{p}_{\mathbf{v}_i}^T \mathbf{p}_{\mathbf{v}_j}$

Let $W_{ij} = \Pr(c_i = j|\mathbf{v}_i)$, learning paradigm:

$$\begin{aligned}&\text{minimize}_W \ ||P - WW^T||_F^2 \\ &\text{subject to} \quad W \in \mathbb{R}_+^{N\times K}, W\mathbf{1}_K = \mathbf{1}_N\end{aligned} \quad \xrightarrow{PenaltyMethod} \quad \begin{aligned}\text{minimize}_W \ &||P - WW^T||_F^2 - \lambda_1 \Sigma_{ij} \min\{0, W_{ij}\} \\ &+ \lambda_2 ||W\mathbf{1}_K - \mathbf{1}_N||_2^2\end{aligned} \quad (1)$$

Sequential minimization: For each fixed pair (λ_1, λ_2) solve the unconstrained problem and then increase (λ_1, λ_2) iteratively until convergence.

	Kernel Kmeans	Spectral Clustering	SymNMF	SoF
Objective	$\min \|\|K - WW^T\|\|_F^2$	$\min \|\|L - WW^T\|\|_F^2$	$\min \|\|A - WW^T\|\|_F^2$	$\min \|\|P - WW^T\|\|_F^2$
Property	K is S.P.D.	L is graph Laplacian, S.P.D.	A is similarity matrix	P is nonnegative, S.P.D.
Constraint	$W^TW = I, W \succeq 0$	$W^TW = I$	$W \succeq 0$	$W \succeq 0, W\mathbf{1} = \mathbf{1}$

Experiments

	Purity							*Rand Index*							*Accuracy*						
	BL	*BT*	*GL*	*IRIS*	*ECO*	*IMG*	*DIG*	*BL*	*BT*	*GL*	*IRIS*	*ECO*	*IMG*	*DIG*	*BL*	*BT*	*GL*	*IRIS*	*ECO*	*IMG*	*DIG*
K-means	0.76	0.43	0.56	0.87	0.80	0.70	0.71	0.60	0.71	0.69	0.86	0.81	0.81	0.91	0.73	0.34	0.51	0.86	0.60	0.63	0.68
S-means	0.76	0.43	0.57	0.92	0.76	0.63	0.71	0.52	0.76	0.69	0.92	0.78	0.81	0.91	0.61	0.39	**0.52**	0.90	0.56	0.61	0.67
GMM	0.77	0.46	0.52	0.89	0.81	0.74	0.70	0.57	0.76	0.64	0.88	**0.85**	0.83	0.91	0.66	0.43	0.46	0.86	0.68	0.69	0.67
SP-NC	**0.78**	0.42	0.38	0.86	0.81	0.30	0.11	**0.64**	0.74	0.43	0.85	0.81	0.32	0.11	0.73	0.41	0.37	0.85	0.61	0.28	0.11
SP-NJW	0.76	0.40	0.38	0.59	0.80	0.48	0.13	0.51	0.71	0.53	0.60	0.80	0.74	0.41	0.56	0.42	0.33	0.56	0.58	0.37	0.12
NMF	0.76	0.43	0.58	0.79	0.73	0.57	0.47	0.56	0.65	0.71	0.81	0.81	0.80	0.86	0.68	0.39	0.50	0.79	0.66	0.53	0.44
RNMF	0.76	0.28	0.52	0.84	0.70	0.56	0.45	0.59	0.38	0.65	0.85	0.78	0.79	0.86	0.71	0.27	0.51	0.84	0.59	0.52	0.41
GNMF	0.76	0.44	0.55	0.85	0.77	0.71	0.71	0.60	0.68	0.70	0.85	0.80	0.77	0.91	0.72	0.36	0.44	0.81	0.56	0.54	0.67
SymNMF	0.76	0.47	0.60	0.88	0.81	**0.77**	0.76	0.62	0.77	0.70	0.88	0.83	0.83	0.91	**0.76**	0.42	0.46	0.87	0.66	0.68	0.67
SoF	0.76	**0.51**	**0.64**	**0.95**	**0.85**	0.75	**0.82**	**0.64**	**0.79**	**0.73**	**0.93**	**0.85**	**0.86**	**0.94**	**0.76**	**0.48**	0.47	**0.94**	**0.74**	**0.71**	**0.82**

a b

[a] Algorithm which achieves best score is highlighted using bold font.
[b] Algorithm which achieves best score and is significantly better than other algorithms under Wilcoxon signed-rank test is highlighted in bold and blue color.

Data sets	# Objects	# Attributes	# Classes
BL	748	4	2
BT	106	9	6
GL	214	9	6
IRIS	150	4	3
ECO	336	7	8
IMG	4435	36	6
DIG	10992	16	10

Analysis

- P nonnegative and S.P.D., an approximation to completely positive matrix.
- The optimization problem is constrained version of completely positive matrix factorization.
- Optimal solution not unique, label switching.

Conclusion

- A novel probabilistic clustering algorithm derived from a set of properties characterizing the co-cluster probability in terms of pairwise distances.
- A relaxation of C.P. programming, revealing a close relationship between probabilistic clustering and symmetric NMF-based algorithms.
- A sequential minimization framework to find a local minimum solution.

References

[1] Dhillon, I. S.; Guan, Y.; and Kulis, B. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowl- edge discovery and data mining.*

[2] Dickinson, P. J., and Gijben, L. On the computational complexity of membership problems for the completely positive cone and its dual. *Computational Optimization and Applications.*

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School of Computer Science