ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

On the equivalence between regularized NMF and similarity-augmented graph partitioning

Anthony Coutant and Hoel Le Capitaine and Philippe Leray

LINA (UMR CNRS 6241) - DUKe Research Group
Ecole Polytechnique de l'Université de Nantes - France

Abstract. Many papers pointed out the interest of (co-)clustering both data and features in a dataset to obtain better performances than methods focused on data only. In addition, recent work have shown that data and features lie in low dimensional manifolds embedded into the original space and this information has been introduced as regularization terms in clustering objectives. Very popular and recent examples are regularized NMF algorithms. However, these techniques have difficulties to avoid local optima and require high computation times, making them inadequate for large scale data. In this paper, we show that NMF with manifolds regularization on a binary matrix is mathematically equivalent to an edge-cut partitioning in a graph augmented with manifolds information in the case of hard co-clustering. Based on these results, we explore experimentally the efficiency of regularized graph partitioning methods for hard co-clustering on more relaxed datasets and show that regularized multi-level graph partitioning is much faster and often find better clustering results than regularized NMF, and other well-known algorithms.

1 Introduction

Non-negative Matrix factorization (NMF) methods are very popular methods which aim at finding a low-rank approximation of a bigger matrix, in order to compress, or cluster data. While many of these techniques make no assumption about the shape of data to find the factored information, recent work [1] have shown that datasets actually lie in a low dimensional manifold embedded in the higher dimensional feature space. Embedding this natural neighborhood relationship between data points has been proposed in recent publications [1, 2] to improve factorizations results. Considering data and features as two independent individual sets, called "modes" of our data, manifold information is encoded as similarity matrices between individuals of same mode, and regularization is done by adding a penalization term involving the Laplacian of these manifolds in the objective function. A common use of NMF is to perform relational clustering between two modes, by factoring the weighted affinity matrix between the two modes sets. Unfortunately, NMF techniques, including the regularized ones, suffer from their scalability and sensitivity to initialization.

Relational clustering can be tackled with different techniques than NMF. We can indeed cast the problem of relational clustering by the one of graph partitioning on the graph induced by the adjacency matrix previously defined. Very efficient and scalable algorithms have been designed for graph partitioning

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

[3], especially multi-level graph partitioning methods (like METIS [4] and GRACLUS [5]). They consist in first coarsening the graph before clustering it, and then uncoarsening data doing local optimization at each step.

In this paper, we show a mathematical equivalence between NMF with manifold regularization for hard clustering of a binary matrix and edge-cut partitioning in the graph built from the union of a relational graph (described by the original matrix) and a set of manifolds graphs (described by similarity matrices). We then explore regularized graph partitioning of less constrained matrices and compare it with several clustering methods, including a state of the art regularized NMF one, and show that regularized multi-level graph partitioning can perform better with a lower computation time than these methods. With the increasing availability of large complex bipartite datasets, our contribution allows relational clustering to scale better with manifolds regularization.

This paper is organized as follows. Section 2 formally describes the problem of relational clustering and demonstrates the equivalence of NMF with manifolds regularization and regularized edge-cut minimization, under hard clustering assumptions. Then, section 3 empirically validates our graph partitioning method over several other clustering algorithms, on 8 UCI datasets.

2 Regularized NMF and regularized graph partitioning

Given two modes, noted $\top$ and $\perp$ and two clusters sets $C_\top$ and $C_\perp$, relational clustering aims at finding two partition functions $P_\top : \top \to C_\top$ and $P_\perp$ defined in a similar way, such that individuals in $c_i \in C_\top$ have close connection pattern with individuals in $\perp$ for all $0 \leq i \leq |C_\top|$ and vice versa. Whenever we also have some extra information on individuals in each set, we can use this information in order to improve the clustering. This is then a regularized relational clustering. In the following, we note $W_\top$ and $W_\perp$ the similarity matrices between individuals of each mode. Similarity matrices can be filled for example by setting 1 between two individuals which are in the k-nearest neighbors set of one another, else 0.

Given a $|\top| \times |\perp|$ matrix A, NMF techniques aims at finding a condensed representation of this original data, by finding low rank matrices $G_\top$ and $G_\perp$, such that $G_\top G_\perp^t$ is a good reconstruction of A, i.e. we have a minimal $d(A, G_\top G_\perp^t)$ for some divergence or distance d. In order to increase the size of the solutions space, and to allow different clusters sets for the two modes, a third factor matrix S is often introduced and the objective is then to find the three matrices such that their product $G_\top S G_\perp^t$ is a good reconstruction of A. A common distance function is the squared Frobenius norm of the difference between A and its reconstruction. Thus, the common objective function of 3-factors NMF is:

$$\Phi = ||A - G_\top S G_\perp^t||_F^2, \; s.t. \; G_\top^t G_\top = I, G_\perp^t G_\perp = I, S > 0 \tag{1}$$

Note that in a clustering context, $G_\top$ and $G_\perp$ represent the assignation of each individual in the corresponding mode to its clusters set, and that in NMF with two reconstruction matrices, we assume $C_\top = C_\perp$.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

In the context where we also have extra information between individuals of each mode, under the form of similarity matrices $W_\top$ and $W_\bot$, we can also add some regularization terms in the objective function. We then have:

$$\Phi_r = ||A - G_\top S G_\bot^t||_F^2 + \lambda_\top tr(G_\top^t L_\top G_\top) + \lambda_\bot tr(G_\bot^t L_\bot G_\bot) \quad (2)$$

where L_i is the Laplacian of W_i and λ_i controls the importance of L_i.

Given a graph $\mathcal{G}(V, E)$, graph partitioning aims at finding a partition of V such that the edge cut is minimal. Denoting by $(V_i)_{0\leq i\leq k}$ a partition of V, we can define the total cut as $\sum_{i<j} w(E_{ij})$ where $E_{ij} \subseteq E$ is the set of edges having one end in V_i and the other in V_j, and w is the function giving the sum of weights of a set of edges. That is, we minimize the weight of edges lying between clusters.

Next, we show that minimizing Φ_r under hard clustering assumption on a binary original matrix is equivalent to performing graph partitioning on the graph $\mathcal{G}(\top \cup \bot, E_r \cup E_\top \cup E_\bot)$, where $E_r \subseteq \top \times \bot$ (resp. $E_\top \subseteq \top \times \top$ and $E_\bot \subseteq \bot \times \bot$), are the inter(resp. intra)-mode sets, with edges weights adjustments.

Theorem 1. *Let $\top$ and $\bot$ be two set of entities and A be the $|\top| \times |\bot|$ affinity matrix between these two sets, with $\forall (i,j) A_{ij} \in \{0, 1\}$. Let also consider intra-modes similarity matrices $W_\top$ and $W_\bot$, their Laplacian $L_\top$ and $L_\bot$ and some non-negative regularization parameters $\lambda_\top$ and $\lambda_\bot$. Assuming we want to find a hard co-clustering of both modes in the same clusters set, then minimizing (2) is equivalent to the edge cut minimization problem in the graph defined by:*

$$X = \begin{bmatrix} \lambda_\top W_\top & A \\ A^t & \lambda_\bot W_\bot \end{bmatrix}.$$

Proof. Considering the NMF objective previously defined, we can write it as:

$$\Phi_r = tr(A^t A + G_\bot S^t G_\top^t G_\top S G_\bot^t - G_\top S G_\bot^t A^t - G_\bot S^t G_\top^t A) + \sum_{i\in\{\top,\bot\}} \lambda_i tr(G_i^t L_i G_i)$$

Since $A^t A$ is a constant, it has no impact on the objective minimization. In addition, using properties of sum and transposes traces, we can write:

$$\arg\min \Phi_r = \arg\min tr(G_\bot S^t G_\top^t G_\top S G_\bot^t) - 2tr(G_\top S G_\bot^t A^t) + \sum_{i\in\{\top,\bot\}} \lambda_i tr(G_i^t L_i G_i)$$

The first term, $tr(G_\bot S^t G_\top^t G_\top S G_\bot^t) = tr((G_\top S G_\bot^t)^t (G_\top S G_\bot^t))$ is the reconstruction matrix norm $||G_\top S G_\bot^t||_F^2$. Thus, the optimization function tends to lower the overall reconstruction matrix values, privileging sparse factors. The two last regularization terms can also be slightly updated by incorporating the regularization parameters into the Laplacian matrices. Thus, we have:

$$\arg\min \Phi_r = \arg\min \quad ||G_\top S G_\bot^t||_F^2 - 2tr(G_\top S G_\bot^t A^t) + \sum_{i\in\{\top,\bot\}} tr(G_i^t \lambda_i L_i G_i)$$

In the case of hard clustering, when G_i matrices describe hard clustering assignments for both modes, the clusters set $C_\top$ and $C_\bot$ have the same cardinalities and S describes a hard assignation of clusters in $C_\top$ to clusters in

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

$C_\perp$ (S is thus a permutation matrix), defining $\widetilde{G}_\perp = G_\perp S^t$, we can rewrite the objective function in a more compact way:

$$\arg\min \Phi_r = \arg\min \quad ||G_\top \widetilde{G}_\perp^t||_F^2 + tr\left(\begin{bmatrix} G_\top^t & \widetilde{G}_\perp^t \end{bmatrix} \begin{bmatrix} \lambda_\top L_\top & -A \\ -A^t & \lambda_\perp L_\perp \end{bmatrix} \begin{bmatrix} G_\top \\ \widetilde{G}_\perp \end{bmatrix}\right) \quad (3)$$

The equivalence holds because we have:

$$tr(G_\perp^t L_\perp G_\perp) = tr(\widetilde{G}_\perp^t L_\perp \widetilde{G}_\perp) = tr(S G_\perp^t L_\perp G_\perp S^t)$$

where S is a permutation matrix. The second term in (3) is close to a Laplacian matrix, the only difference being in the diagonal, since the regularization Laplacian matrices only involve the intra-mode degrees. However, since the inter-mode degrees are constant given A, the minimization problem does not change when adding the inter-mode degrees term. Finally, we can write:

$$\arg\min \Phi_r = \arg\min ||G_\top \widetilde{G}_\perp^t||_F^2 + tr\left(\begin{bmatrix} G_\top^t & \widetilde{G}_\perp^t \end{bmatrix} (D - X) \begin{bmatrix} G_\top \\ \widetilde{G}_\perp \end{bmatrix}\right)$$

with D the degree matrix of X, and $D-X$ the Laplacian L of X. The objective is thus equivalent to an edge cut minimization in X, plus an extra term privileging sparse solutions, under our assumptions, which concludes the proof. □

3 Experiments

Note that even if objective functions have been proved equivalent in last section under our assumptions, the obtained results in practice are sensitive to the algorithms heuristics, which are different. Thus, it may not be possible to observe a strict equivalence in practice, even if we strictly reproduce our assumptions.

We use our theoretical results as a start to explore the efficiency of regularized graph partitioning for relational clustering. In practice, the binary matrix, permutation matrix S and same clusters sets assumptions are not always realistic, thus we make our experiment in a relaxed context. More precisely, we compare the performances and running times of six algorithms, including two graph partitioning ones, on 8 UCI [6] datasets: *Glass* (214 data, 9 features, 6 clusters to find), *Heart* (270, 13, 2), *Semeion* (1593, 256, 10), *Soybean* (47, 35, 4), *SPECTF* (267, 45, 2), *Vehicle* (846, 19, 4), *Wine* (198, 33, 2).

Compared algorithms are: 1) k-means; 2) projected gradient NMF [7]; 3) normalized cut (NCut) [8], using a nearest-neighbors approach for the affinity matrix (only the data manifold is considered); 4) graph dual regularization non-negative matrix tri-factorization (DNMTF) [2], a state of the art algorithm for regularized NMF, where the manifolds matrices are built using the nearest-neighbors approach; 5) METIS graph partitioning with manifolds regularization (rMETIS); 6) GRACLUS graph partitioning with manifolds regularization (rGRACLUS). Following results of section 2, we expect the regularized graph partitioning algorithms to perform closely to DNMTF, with lower running times.

Every algorithm concerned by manifold regularization (i.e. NCut, DNMTF, rMETIS, rGRACLUS) defines its similarity (or affinity for NClust) matrices $W_\top$

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Data	Eval.	K-means	NMF	NCut	DNMTF	rMET	rGRA
Glass	NMI	340±4	270±60	339±2	335±21	**407±0**	395±0
	ACC	536±15	514±36	625±2	592±25	**696**±0	**696±0**
	Time	10±0	850±170	70±0	90±20	50±0	30±0
Heart	NMI	217±96	70±18	100±0	217±39	**323±0**	234±0
	ACC	753±80	654±16	652±0	754±34	**819±0**	778±0
	Time	0±0	140±50	160±10	570±0	30±0	30±0
Semeion	NMI	540±10	334±11	604±5	377±6	664±0	**709**±0
	ACC	627±10	409±16	636±0	458±25	747±0	**788**±0
	Time	390±120	3560±1130	1910±100	53560±460	290±10	190±10
Soybean	NMI	879±110	888±33	**1000±0**	954±59	**1000±0**	**1000±0**
	ACC	915±81	940±28	**1000±0**	974±34	**1000±0**	**1000±0**
	Time	0±0	180±80	30±0	120±30	40±0	30±0
SPECTF	NMI	78±7	94±20	117±0	92±14	165±0	**173±0**
	ACC	794±0	794±0	794±0	**795±1**	794±0	794±0
	Time	0±0	140±30	70±10	900±0	40±0	50±0
Vehicle	NMI	101±0	33±10	159±0	136±24	245±0	**250±0**
	ACC	405±1	330±13	447±0	429±34	**515±0**	508±0
	Time	20±10	430±70	170±0	4000±930	50±0	50±0
Wine	NMI	837±27	697±25	907±0	722±31	895±0	**921±0**
	ACC	951±11	892±20	978±0	908±14	966±0	**978±0**
	Time	0±0	100±10	30±0	280±160	40±0	30±10
Wpbc	NMI	19±6	23±4	31±1	47±8	**60±0**	57±0
	ACC	763±0	763±0	763±0	**771±5**	763±0	768±0
	Time	0±0	70±10	40±0	590±0	40±10	30±0

Table 1: NMI and ACC best mean evaluations + standard deviations ($\times 10^{-3}$). Corresponding running times are given (in ms.).

and $W_\perp$ by choosing a number $n_\top$ of nearest neighbors for each data individual, and $n_\perp$ for each feature vector, and assigning a 1 to every element of the matrices relating nearest neighbors to each other, else 0. Remind that the two graph partitioning methods are performed on an augmented graph adding intra-mode edges weighted by these similarities. $n_\top$ and $n_\perp$ are each taken in the grid $\{1, 2, 3, 4, 5, 6, 7, 9, 11\}$ and regularization parameters $\lambda_\top$ and $\lambda_\perp$ are each taken in the grid $\{1, 10, 100, 500\}$. Every parameters combination is computed 5 times.

Algorithms are evaluated by their Normalized Mutual Information (NMI) and clustering accuracy (ACC) mean over the 5 iterations, on every dataset normalized by column, given the right number of clusters. Best means, corresponding standard deviations and running times are given in Table 1 for every (dataset, algorithm, evaluation measure) triple.

We can see that regularized METIS and GRACLUS achieves good performances. Surprisingly, they mainly succeed in achieving the best performances, both for NMI and ACC. This can be explained by the high variance of DNMTF, even given a set of parameters, since we evaluate the best mean over several folds for these criteria. In some other cases whenever DNMTF achieves the best performances, graph partitioning algorithms are still close. We can

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

also see that performances of NCut are good but most often below ones of regularized partitioning methods. This is interesting, since NCut only uses one mode neighborhood information. Thus, it seems that the use of both manifolds regularization and the relational data is profitable.

In terms of execution times, DNMTF is significantly more time consuming than graph partitioning algorithms, and its performances grow significantly with matrix size. As an example, the average time on Semeion dataset is around 50 sec. for DNMTF, against 0.33 sec. for GRACLUS and METIS.

4 Conclusion

In this paper, we first have demonstrated that NMF with manifolds regularization on a binary matrix is mathematically equivalent to an edge-cut partitioning in a manifold augmented graph, for hard co-clustering task. We have then shown experimentally the efficiency of regularized graph partitioning methods for hard co-clustering, both on binary and more relaxed datasets, in terms of clustering performance and time. Perspectives for this work include the exploration of more theoretical aspects for the equivalence between regularized NMF and regularized graph partitioning, especially when relaxing our assumptions. In addition, other regularizations have to be investigated, like similarities from attributes values.

References

[1] Deng Cai, X He, X Wu, and J Han. Non-negative matrix factorization on manifold. In *ICDM*, pages 63–72, 2008.

[2] Fanhua Shang, LC Jiao, and Fei Wang. Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition*, 45(6):2237–2250, June 2012.

[3] Aydin Buluç, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. Recent Advances in Graph Partitioning. In *Algorithm Engineering - Selected Topics, to app., ArXiv:1311.3144, 2014.*

[4] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.

[5] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Weighted Graph Cuts without Eigenvectors : A Multilevel Approach. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 29(11):1944–1957, 2007.

[6] Kevin Bache and Moshe Lichman. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[7] Chih-jen Lin. Projected Gradient Methods for Non-negative Matrix Factorization. *Neural computation*, 19(10):2756–2779, 2007.

[8] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.