A Unified Non-Negative Matrix Factorization Framework for Semi Supervised Learning on Graphs

Anasua Mitra

Created on 10/02/2020

Semi Supervised Learning (SSL)

— For learning a meaningful inference[1], a data point x should carry useful information for estimating the target function y, i.e., $\Pr(x)$ should help inferring $\Pr(y|x)$.

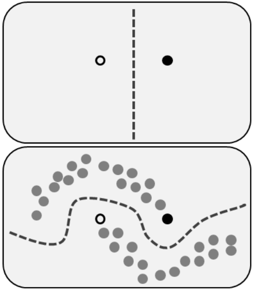


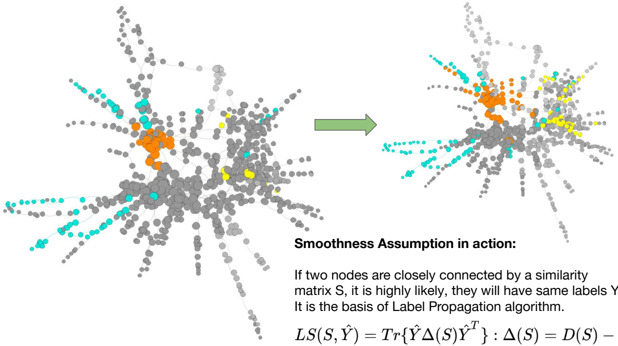
Figure: The influence of unlabeled data in semi-supervised learning (Source: Wikipedia)

- Learning from both **Labeled and Unlabeled data**.
- Unlabeled data is **abundantly available**, unlike costly labeled data!
- Unlabeled data *can* give a **better** sense of class separation boundary!
- **Important prerequisite** – certain assumptions need to be hold.

[1] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)". In: IEEE Transactions on Neural Networks (2009).

Assumptions of SSL

Smoothness/ Continuity Assumption — Enforced between a pair of points[2].



Smoothness Assumption in action:

If two nodes are closely connected by a similarity matrix S, it is highly likely, they will have same labels Y. It is the basis of Label Propagation algorithm.

$$LS(S, \hat{Y}) = Tr\{\hat{Y}\Delta(S)\hat{Y}^T\} : \Delta(S) = D(S) - S$$

Smoothness Assumption

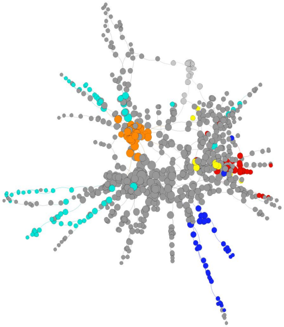
The target function of two closely connected points **in a dense region** should also be close.

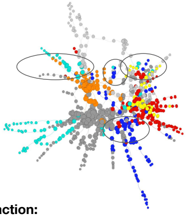
Similar to Supervised Learning assumptions, in addition to that, SSL takes the density of data points into account.

[2] Chapelle, Scholkopf, and Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)".

Assumptions of SSL

Cluster Assumption — A special form of Continuity, enforced among a group of points[3].





Cluster Assumption in action:

If two nodes are connected by paths in dense region, it is highly likely, they will have same labels. Because examples of same class data tend to form clusters.

If E is node-to-node label similarity and H is a cluster matrix that we want to learn based on label invariance,

$$LS(E,H) = Tr((H)\Delta(E)(H)^T) : \Delta(E) = D(E) - E$$

Cluster Assumption

Points belonging to the same cluster are likely to be of the same class as **data from each class follows a coherent distribution, tends to form clusters.**

Although, data that shares a label may spread across multiple clusters.

[3] Olivier Chapelle, Jason Weston, and Bernhard Schölkopf. "Cluster kernels for semi-supervised learning". In: **Advances in neural information processing systems**. 2003.

Assumptions of SSL

Low Density Separation — A preference for decision boundaries in low-density regions[4].

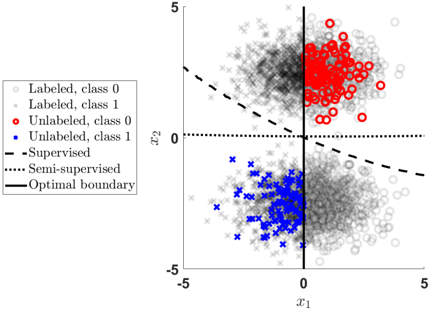


Figure: Decision boundaries of learning algorithms.
Source: Google.

Low Density Separation

The decision boundary should lie in a low-density region.

Continuity Assumptions imply this.

[4]Chapelle, Scholkopf, and Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)".

Assumptions of SSL

Manifold Assumption — To mitigate the *Curse of Dimensionality*.

Manifold Assumption

The data lie approximately on a manifold of much lower dimension than the input space.

Facilitates learning using distances and densities defined on the manifold.

The classical graph based semi-supervised learning loss function can be written as[a],

$$\sum_{i=1}^{L} l\{Y_i, f(X_i)\} + \lambda . \sum_{i,j} A_{i,j} \{f(X_i) - f(X_j)\}^2 \quad (1)$$

Where, X is either original network features, or a low dimensional representation of nodes — which NRL methods facilitates by learning an intermediate function, $g : U \mapsto X$ to project underlying graph's large feature space U to a lower dimensional manifold $X : X \ll U$. $f : X \mapsto Y$ — is a function that predicts a node's labels.

[a] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. "Revisiting Semi-Supervised Learning with Graph Embeddings". In: **International Conference on Machine Learning**. 2016.

Our Contributions

— Encoding SSL Cluster Assumption.

USS-NMF: **Encoding** largely ignored **cluster assumption** to learn clusterable representations of nodes in a transductive graph based SSL framework. We propose **Semi-Supervised Cluster Invariance** Property for nodes, for clustering nodes with similar labels together. **We provide a framework which incorporates essential learning principles of SSL.**

The primary distinction between other graph based SSL methods and ours lies in the fact that,

- We learn a function $h : X \mapsto H$ that learns a node's cluster structure, one abstract space. It is learned along with the label prediction function f to predict labels Y.
- We enforce label invariance, i.e., two nodes with same labels should belong to same clusters, in terms of a train-label similarity matrix E on the cluster space H via Laplacian regularization objective.

Proposed Method: USS-NMF

— Encoding local invariance or network structure[5]. Via network-proximity matrix factorization.

Enforces Manifold Assumption via low-dimensional network representation learning.

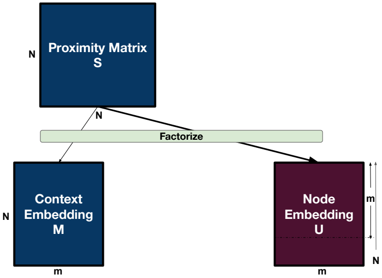


Figure: Encoding Network Structure

$$O_{network} = \min_{M,U} \|\mathbf{S} - \mathbf{U^T M}\|^{\mathbf{2}} : M \geq 0, U \geq 0 \tag{2}$$

[5] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations". In: **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining**. 2014.

Proposed Method: USS-NMF

— Encoding supervision knowledge[6]. Via label matrix factorization & label propagation objectives.

Label Smoothness Assumption enforcing Low-Density Separation.

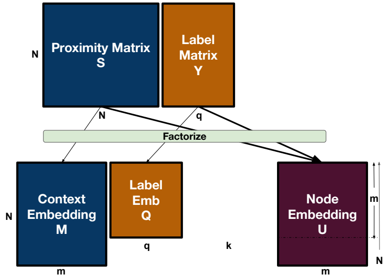


Figure: Encoding Supervision Knowledge

$$O_{label} = \min_{Q,U} \|\mathbf{W} \odot (\mathbf{Y} - \mathbf{QU})\|^{\mathbf{2}} + \mathbf{Tr}\{(\mathbf{QU})\delta(\mathbf{S})(\mathbf{QU})^{\mathbf{T}}\} : Q, U \geq 0 \quad (3)$$

[6] Cunchao Tu et al. "Max-margin deepwalk: Discriminative learning of network representation.". In: IJCAI. 2016.

Proposed Method: USS-NMF

— **Encoding semi-supervised cluster structure**[7]**.** Via cluster membership matrix learning & factorization.

Semi-Supervised Cluster Assumption enforcing Low-Density Separation.

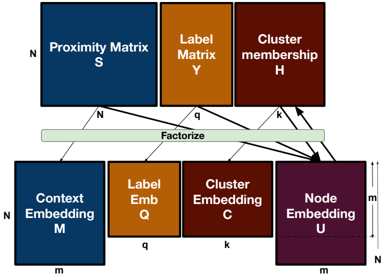


Figure: Encoding Cluster Structure

$$O_{cluster} = \min_{H,C,U \geq 0} \|\mathbf{H} - \mathbf{CU}\|^{\mathbf{2}} + \|\mathbf{HH^T} - \mathbf{I}\|^{\mathbf{2}} + \mathbf{Tr}((\mathbf{H})\delta(\mathbf{E})(\mathbf{H})^{\mathbf{T}}) \tag{4}$$

[7] Xiao Wang et al. "Community preserving network embedding". In: **Thirty-first AAAI conference on artificial intelligence**. 2017.

Proposed Method: USS-NMF

— Encoding semi-supervised cluster structure[8] Via cluster membership matrix learning & factorization.

Semi-Supervised Cluster Assumption enforcing Low-Density Separation.

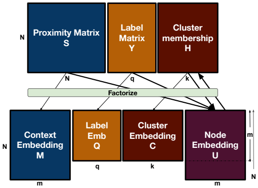


Figure: Encoding Cluster Structure

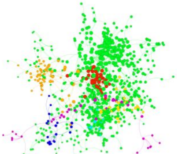
$$O_{cluster} = \min_{H,C,U\geq 0} ||\mathbf{H} - \mathbf{CU}||^2 + ||\mathbf{HH^T} - \mathbf{I}||^2 + \mathbf{Tr}((\mathbf{H})\delta(\mathbf{E})(\mathbf{H})^\mathbf{T})$$

- Cluster structure and nodes influence each other via joint factorization and learning.
- Optimizing un-even distinct cluster membership learning for each node.
- Label invariance in cluster similarity kernel for cluster learning.

[8] Wang et al., "Community preserving network embedding".

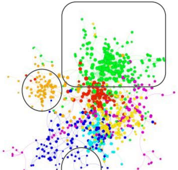
Interesting Visualizations

Q. How competitively the algorithms perform?— Visualizations: on test data.



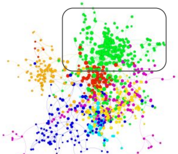
Error due to not preserving label/ topological information

MFDW (Acc: 75.94%)



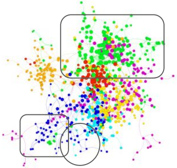
Error due to only label smoothening

MFDW+Y (Acc: 85.17%)



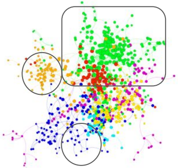
Error due to only unsupervised clustering

MNMF (Acc: 81.99%)



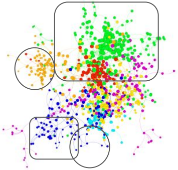
Error due to conflicting joint optimization

MNMF+Y (Acc: 82.51%)



Smooths on label and has more label diversity due to cluster structure

USS-NMF (Acc: 86.72%)



Original

State-Of-The-Art Results

— On node classification & clustering.

- **Cluster enforcing models are the superior unsupervised models.** Among them, Com-E[9], M-NMF[10] showed superior performance than GEMSEC[11].
- All supervised models obtain better performance over unsupervised counter-parts.
- Our model outperformed present SOTA unsupervised community enhanced NRL algorithms M-NMF, COM-E, GEMSEC by a large margin — which shows the **superiority of semi-supervised clustering criteria over any kind of unsupervised clustering criteria**.
- Experiments results for USS-NMF :
  - **Robust performance** (ranks first in 12/13 datasets and ranks second in just 1) across all 13 datasets in comparison with 8 baselines for node classification.
  - **Performs outstandingly well in node clustering** task with improvement upto 7% on average over the second best model MNMFL.
  - **Well-separated & homophilous clusters** obtained in t-SNE visualizations.
  - **USS-NMF does well in both random and balanced test-train splits, even in label sparsity!**, outperformed Planetoid-G[12] by a large margin in their balanced sampling based test-train splits.

[9] Sandro Cavallari et al. "Learning community embedding with community detection and node embedding on graphs". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017.

[10] Wang et al., "Community preserving network embedding".

[11] Benedek Rozemberczki et al. "Gemsec: Graph embedding with self clustering". In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2019.

[12] Yang, Cohen, and Salakhudinov, "Revisiting Semi-Supervised Learning with Graph Embeddings".

Interesting Visualizations

Q. To cluster or not to cluster?— Visualizations & parameter sensitivity analysis.

— USS-NMF provides **well separable homophilous clusters**. Learning clusters almost always improves performance.

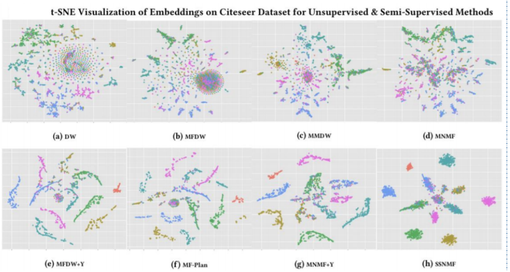


Figure: t-SNE Visualizations

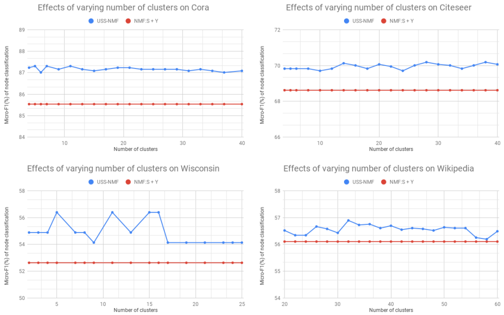


Figure: Varying Number of Clusters

Interesting Visualizations

Q. To cluster or not to cluster? — Ablation study.

— **It is important to analyze the importance of utilizing label and cluster information, separately.**
The figures below show the contribution of label information (left) and cluster information (right), as well as, contributions of their various components. **The boxplots depict statistics of performance improvement (minimum, maximum, mean values, all the quartiles) across all the datasets** owing to each component, separately & collectively — over Matrix Factorized DeepWalk.

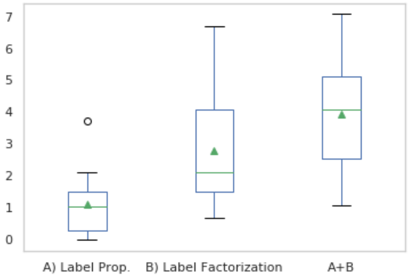


Figure: Usefulness of label information

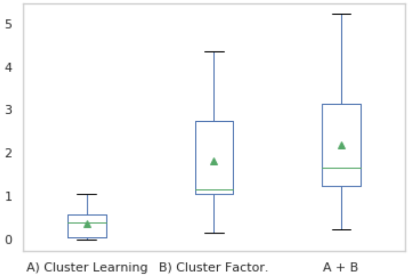


Figure: Usefulness of cluster information

Useful Insights

Q. What about large hyper-parameter search space & performance in *label sparsity*?

— We provide you with one **effective range of hyper-parameter space**. A range, which also gave **decent performance in label sparsity**.

Co-efficients	**USS-NMF** (Effective range)	
Dataset (small=$<$1k, large$>$1k —V—)	Small	Large
Network	1, 5	1, 5, 10
Label	0.1, 1	0.1, 0.5, 1
Cluster Factorization	0.1, 1	0.1, 0.5, 1
Cluster Learning	10	10
Cluster Orthogonality	1e + (0, 4)	1e + 8
Graph Laplacian Regularization	0.5, 1	0.5, 1
L2 Regularization	1	1
#Clusters	#Labels	#Labels
#Experiments	32 (Full search)	54 (Full search)

Table: Hyper-parameter search space for USS-NMF

Dataset	**Cora**					
Train (%)	5	10	20	30	40	50
NMF:S+Y	67.519	76.620	79.012	84.660	84.194	85.535
MMDW	67.403	75.101	80.093	82.942	82.889	83.838
MNMF+Y	68.947	76.948	81.172	84.396	83.518	85.904
USS-NMF	**69.452**	**78.015**	**82.880**	**85.609**	**85.486**	**87.380**

Table: Node Classification Results — Varying Train-Test Splits — Micro-F1 Scores

Thank You