Réduction de dimension et classification non supervisée d’un graphe attribué : une approche simultanée

Lazhar Labiod and Mohamed Nadif

LIPADE, University of Paris, 45 rue des Saints-Pères, Paris, France.

04/09/2019
Plan

1. Introduction
2. Attributed network embedding and clustering: a tandem approach
   - Spectral clustering
   - AN Embedding and clustering: a tandem approach
3. Simultaneous attributed network embedding and clustering
   - SANEC: Data integration
   - SANEC approach
4. Experiments
   - Clustering Results
   - Data visualization
5. Conclusion
Attributed Network

- AN are used to model a large variety of real-world networks, such as academic networks and health care systems where both node links and attributes/features are available for analysis.

- AN embedding and clustering has received a significant amount of attention as an important problem with many applications including social networks, academic citation networks and protein-protein interaction networks.

- Unlike plain networks in which only node links and dependencies are observed, each node in an AN is often associated with a valuable set of features.
Attributed Network(AN)

**Definition**

An attributed network $G = (V, E, X)$ consists of $V$ the set of nodes, $E \subseteq V \times V$ the set of links, and $X = [x_1, x_2, \ldots, x_n]$ where $n = |V|$ and $x_i \in \mathbb{R}^d$ is the feature/attribute vector of the node $v_i$.

**Attributed Network**

Formally, the graph can be represented by two types of information;

- the content information $X \in \mathbb{R}^{n \times d}$.
- the structure information $A \in \mathbb{R}^{n \times n}$, where $A$ is an adjacency matrix of $G$ and $a_{ij} = 1$ if $e_{ij} \in E$ otherwise 0; we consider that each node is a neighbor of itself, then we set $a_{ii} = 1$ for all nodes.
AN embedding and clustering

Clustering

The aim of cluster analysis is the discovery of a finite number of homogeneous classes from data.

Embedding

AN embedding aims to seek a continuous low-dimensional matrix representation for nodes in a network, such that original network topological structure and node attribute proximity can be preserved in this new low-dimensional embedding.

Embedding and clustering

- PCA + k-means.
- Spectral embedding + k-means.
Spectral Clustering (SC) [NJW, 2001]

- **Step 1.** finding a spectral embedding by using an eigenvector/eigenvalue decomposition of a Laplacian matrix.
- **Step 2.** K-means clustering is performed on the first few eigenvectors.

**Remark**

In Spectral Clustering, spectral embedding (Step 1) and clustering (Step 2) tasks are performed separately.
Graph Embedding and clustering: a tandem approach

**Figure** - Spectral Clustering.
AN Embedding and clustering: a tandem approach

Figure — AN Embedding.
Although existing AN clustering has been applied in practice widely, they may easily achieve poor performance due to the following drawbacks:

- The disadvantage of this approach is that consists in optimizing two different objectives.
- High risk of severe deviation of approximate continuous embedding solution from the good discrete clustering.
- Information loss among separate independent stages, i.e., continuous embedding generation, embedding discretization.
SANEC

We propose a novel simultaneous AN Embedding and clustering scheme which jointly:

- learns embedding from both network and attributes information.
- learns continuous embedding and discrete clustering labels.
- Specifically, we explicitly enforce a discrete transformation on the intermediate continuous labels (embedding).
- This leads to a tractable optimization problem with a discrete solution.
SANEC framework

**Figure** - SANEC Framework
**SANEC framework**

**W, M and S**

- **Nodes proximity**: \( W = D^{-1}A \), where \( D \) is the degree matrix of \( A \) defined by \( d_{ii} = \sum_{i'}=1 a_{i'i} \).

- **New data representation**: \( M = (m_{ij}) \) can be considered as a multiplicative integration of both \( W \) and \( X \) by replacing each node by the centroid of their neighborhood (barycenter): i.e., \( m_{ij} = \sum_{k=1}^{n} w_{ik} x_{kj} \), \( \forall i, j \) or \( M = WX \).

- **Additional information about nodes similarity from X**: we choose to take \( S \) defined by \( S = W + WX \).
SANEC Objective Function

\[
\min_{B, Z, Q, G} \left\| M - BQ^T \right\|^2 + \lambda \left\| S - GZB^T \right\|^2
\]

\[B^T B = I, \ Z^T Z = I, \ G \in \{0, 1\}^{n \times k}\]

- \(G = (g_{ij})\) of size \((n \times k)\) is a cluster membership matrix,
- \(B = (B_{ij})\) of size \((n \times k)\) is the embedding matrix;
- \(Z = (q_{ij})\) of size \((k \times k)\) is an orthonormal rotation matrix which most closely maps \(B\) to \(G\).

Spectral Embedding criterion

\[
\min_{B, Q} \left\| M - BQ^T \right\|^2.
\]

Spectral rotation (in low embedding subspace)

\[
\left\| S - GZB^T \right\|^2 = \left\| S - BB^T S \right\|^2 + \|SB - GZ\|^2
\]
SANEC : Optimization

To solve (1) we rely on the following theorem.

**Theorem**

Let $X_{n \times k}$ and $B_{n \times k}$ be two matrices. Consider the constrained minimization problem

$$Z_* = \arg \min_Z \left\| X - BZ^T \right\|^2 \quad s.t \quad Z^T Z = I.$$

Suppose the SVD of $X^T B$ is $UDV^T$, then $Z_* = UV^T$.

We can observe that this problem turns out to be similar to the well known orthogonal Procrustes problem (Schonemann, 1966)
SANEC : Optimization

**Compute Z.** Fixing $G$ and $B$ the problem which arises in (1) is equivalent to

$$\max_Z \text{Tr}(G^\top SBZ) \quad \text{s.t.} \quad Z^\top Z = I. \quad (2)$$

**Compute Q.** Given $G$, $Z$ and $B$, the optimization problem (1) is equivalent to

$$\min_Q \|M - BQ^\top\|^2. \quad (3)$$

**Compute B.** Given $G$, $Q$ and $Z$, the problem (1) is equivalent to

$$\max_B \text{Tr}((M^\top Q + \lambda SGZ)B^\top) \quad \text{s.t.} \quad B^\top B = I. \quad (4)$$

**Computation of G.** As $G$ is a cluster membership matrix, its computation is done as follow : We fix $Q$, $Z$, $B$. Let $\tilde{B} = SB$ and calculate

$$g_{ik} = \begin{cases} 
1 & \text{if } k = \arg \min_{k'} \|\tilde{b}_i - z_{k'}\|^2 \\
0 & \text{otherwise}.
\end{cases} \quad (5)$$
In summary, the steps of the SDRC algorithm can be deduced in Algorithm 1.

**Algorithme 1 : SANEC algorithm**

- **Input**: $M$ and $S$ from structure matrix $W$ and content matrix $X$;
- **Initialize**: $B$, $Q$ and $Z$ with arbitrary orthonormal matrix;
- **repeat**
  - (a) - Compute $G$ using (5)
  - (b) - Compute $B$ using (4)
  - (c) - Compute $Q$ using (3)
  - (d) - Compute $Z$ using (2)
- **until** convergence
- **Output**: $G$ : clustering matrix, $Z$ : rotation matrix, $B$ : nodes embedding and $Q$ : attributes embedding
Table – Description of Data sets

<table>
<thead>
<tr>
<th>Data sets</th>
<th>#Nodes</th>
<th># Attributes</th>
<th>#Edges</th>
<th>#classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cora</td>
<td>2708</td>
<td>1433</td>
<td>5294</td>
<td>7</td>
</tr>
<tr>
<td>Citeseer</td>
<td>3312</td>
<td>3703</td>
<td>4732</td>
<td>6</td>
</tr>
<tr>
<td>Wiki</td>
<td>2405</td>
<td>4973</td>
<td>17981</td>
<td>19</td>
</tr>
</tbody>
</table>

- **Cora**: A citation network with 2708 nodes and 5294 links between them; the nodes correspond to publications described by binary vectors of 1433 dimensions and classified into 7.
- **Citeseer**: A citation network consisting of 3312 publications labeled into 6 sub-fields. Each publication is described by a binary vector of 3703 dimensions and there are 4732 links between them.
- **Wiki**: A network with 2405 documents and 17981 links. These documents have 4973-dimension vectors representing them and are divided into 19 classes.
Compared algorithms

- Graph Encoder [16] learns graph embedding for spectral graph clustering.
- RTM [3] learns the topic distributions of each document from both text and citation.
- DeepWalk[10] is a network representation approach which encodes social relations into a continuous embedding space.
- Spectral Clustering[14] is a widely used approach for learning social embedding.
- GAE [8] is an autoencoder-based unsupervised framework for attributed network data embedding.
- VGAE [8] is a variational graph autoencoder approach for graph embedding with both node links and node attributes information.
- ARGA is the most recent adversarially regularized autoencoder algorithm which uses graph autoencoder to learn the embedding.
- ARVGA [9] algorithm, which uses a variational graph autoencoder to learn the embedding.
### Clustering Results

**Table** – Clustering performances (Acc %, NMI % and ARI %) on Cora, Citeseer and Wiki datasets

| Methods        | Datasets          | Cora | | | | | | Citeseer | | | | | | Wiki | | | | |
|----------------|-------------------|------|---|---|---|---|---|------|---|---|---|---|------|---|---|---|---|
|                |                   | Acc  | NMI | ARI | Acc  | NMI | ARI | Acc  | NMI | ARI | Acc  | NMI | ARI  |
| K-means        |                   | 49.22| 32.10| 22.96| 54.01| 30.54| 27.86| 41.72| 44.02| 15.07|
| Spectral       |                   | 36.72| 12.67| 03.11| 23.89| 05.57| 01.00| 22.04| 18.17| 01.46|
| GraphEncoder   |                   | 32.49| 10.93| 00.55| 22.52| 03.30| 01.00| 20.67| 12.07| 0.49 |
| DeepWalk       |                   | 48.40| 32.70| 24.27| 33.65| 08.78| 09.22| 38.46| 32.38| 17.03|
| DNGR           |                   | 41.91| 31.84| 14.22| 32.59| 18.02| 04.29| 37.58| 35.85| 17.97|
| RTM            |                   | 43.96| 23.01| 16.91| 45.09| 23.93| 20.26| 43.64| 44.95| 13.84|
| RMSC           |                   | 40.66| 25.51| 08.95| 29.50| 13.87| 04.88| 39.76| 41.50| 11.16|
| TAWD           |                   | 56.03| 44.11| 33.20| 45.48| 29.14| 22.81| 30.96| 27.13| 04.54|
| VGAE           |                   | 50.20| 32.92| 25.47| 46.70| 26.05| 20.56| 45.09| 46.76| 26.34|
| ARGE           |                   | 64.00| 44.90| 35.20| 57.30| 35.00| 34.10| 47.34| 47.02| 28.16|
| ARVGE          |                   | 63.80| 45.00| 37.74| 54.40| 26.10| 24.50| 46.45| 47.80| 29.65|
| SANEC          |                   | **67.38**| **47.14**| **39.88**| **66.77**| **40.60**| **41.78**| **49.57**| **48.30**| **33.14**|
Cora visualization

**Figure** – Cora visualization.
Visualization of Citeseer dataset

**Figure** – Visualization of Citeseer dataset.
Sensitivity analysis according to $\lambda$

**Figure** – Sensitivity analysis.
### Contributions

- We proposed a novel matrix decomposition framework for simultaneous attributed network data embedding and clustering.

- We show that the objective learning of SANEC can be decomposed into three terms, the first one is the objective function of PCA applied to $M = WX$, the second is the graph embedding criterion in a low-dimensional space, and the third term is the clustering criterion.

- We also integrate a discrete rotation functionality, which allows a smooth transformation from the relaxed continuous embedding to a discrete solution and guarantees a tractable optimization problem with a discrete solution.
Thank You