KDD 2020 paper reading 10/15

Houquan Zhou

October 15, 2020

Information

- **Title**: Data Compression as a Comprehensive Framework for Graph Drawing and Representation Learning
- **Authors**: Claudia Plant, Sonja Biedermann, Christian Böhm
- **Affliations**: Univerity of Vienna, LMU Munich

Data Compression as a Comprehensive Framework for Graph Drawing and Representation Learning

Claudia Plant
Faculty of Computer Science, ds:UniVie
University of Vienna, Vienna, Austria
claudia.plant@univie.ac.at

Sonja Biedermann
Faculty of Computer Science
University of Vienna, Vienna, Austria
sonja.biedermann@univie.ac.at

Christian Böhm
Institute of Informatics, MCML
LMU Munich, Germany
boehm@ifi.lmu.de

Overview

- **Problem**: Representation learning.
- **Idea**: A novel objective function from a data compression perspective: **Predictive Entropy (PE)**.
- **Algorithm**: A novel representation learning algorithm GEMPE.
- **Experiments**: Graph drawing & Reprentation learning.

Predictive Entropy

Embedings quality: The more effectively the low-dimensional embeddings allow to compress the adjacency matrix, the better is the quality of the embedding.

Based on the idea, the authors proposed a novel information-theoretic evaluation function: Predictive Entropy (PE).

PE is based on a probabilistic model, which predicts the probability $\Pr((i,j) \in E)$ from their embeddings x_i, x_j.

Ideal case: $\Pr((i,j) \in E)$ close to 1 for $(i,j) \in E$, and close to 0 for $(i,j) \notin E$.

Probalistic model

Edge probability of (i, j) related to the Euclidean distance of x_i and x_j. The smaller $\|x_i - x_j\|$ is, the more possible i and j are connected.

$$s^+_{\mu,\sigma}(\delta) \coloneqq P\left((i,j) \in E \mid \|x_i - x_j\| = \delta\right) = \frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(\frac{\delta - \mu}{\sqrt{2}\sigma}\right) \quad (1)$$

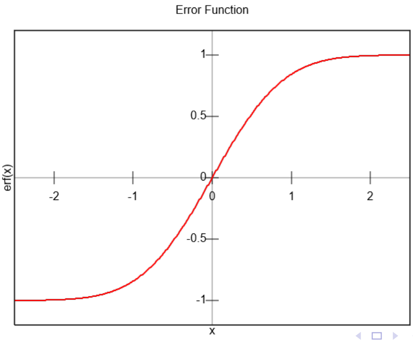
$$s^-_{\mu,\sigma}(\delta) \coloneqq P\left((i,j) \notin E \mid \|x_i - x_j\| = \delta\right) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{\delta - \mu}{\sqrt{2}\sigma}\right) \quad (2)$$

erf function

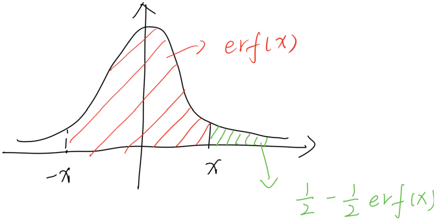
$$\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \mathrm{d}t = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} \mathrm{d}t \tag{3}$$

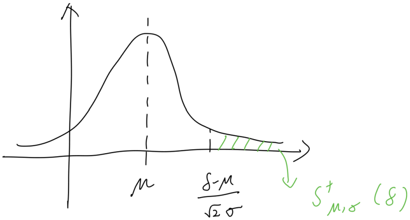
Related to cumulative probability function of Gaussian distribution:

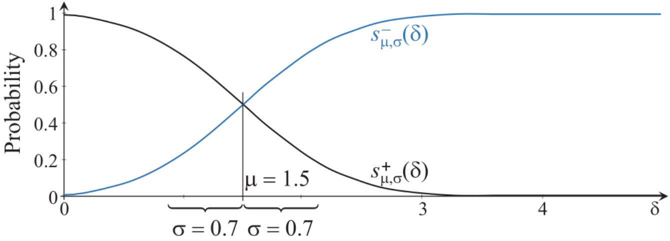
$$\Phi(x) = \frac{1}{2} + \frac{1}{2} \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) \tag{4}$$



Demonstration







Predictive Entropy

Encoding length of (i, j) entry of the adjacency matrix:

$$dl_{\mu,\sigma}(i,j) = \begin{cases} dl^{+}_{\mu,\sigma}(\delta) & \text{if } (i,j) \in E \\ dl^{-}_{\mu,\sigma}(\delta) & \text{otherwise} \end{cases}$$

$$dl^{+}_{\mu,\sigma}(\delta) = -\log\left(\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(\frac{\delta-\mu}{\sqrt{2}\sigma}\right)\right)$$

$$dl^{-}_{\mu,\sigma}(\delta) = -\log\left(\frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{\delta-\mu}{\sqrt{2}\sigma}\right)\right)$$

Total objective function PE:

$$PE_G(x_1, \ldots, x_n) = \frac{1}{N} \min_{\mu,\sigma} \sum_{1 \le i \le n} \sum_{i < j \le n} dl_{\mu,\sigma}(i,j)\,, \tag{5}$$

where $N = \frac{n(n+1)}{2}$, $x_1, \ldots, x_n$ are embeddings of nodes.

Summary: PE

$$
\begin{tikzcd}
G \arrow[r, "\text{repr. learning}"] \arrow[rr, bend right, no head, "\text{PE}"'] & \text{embeddings} \arrow[r, "\text{probablistic model}"] & G'
\end{tikzcd}
$$

- Learn embedding from G.
- Use embeddings to predict the edges.
- Encoding cost $\Rightarrow$ PE.

Weighted Majorization[1]

Definition (Weighted Majorization)

- Given n objectes, distance matrix $d_{i,j}$ and weights $w_{i,j}$;
- Find embeddings $x_i, i = 1, 2, \ldots, n$;
- To minimize the following objective:

$$wm(x_1, \ldots, x_n) = \sum_{1 \leq i \leq n} \sum_{i < j \leq n} w_{i,j}(\|x_i - x_j\| - d_{i,j})^2 \tag{6}$$

WM aims at finding low-dimensional coordinates x_i that minimize the squared difference between the distance in the input matrix $d_{i,j}$ and the Euclidean distance $\|x_i - x_j\|$ of the generated vectors.

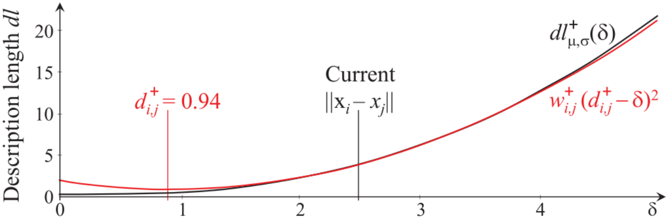
[1] Brandes and Pich 2009; Gansner, Koren, and North 2004.

Optimizing PE using WM

Focusing on one pair (i, j): $w_{i,j}(\|x_i - x_j\| - d_{i,j})^2$.

We want the WM objective function to approximate our objective function. $\Rightarrow$ Two-order approximation.

$$\frac{\mathrm{d}}{\mathrm{d}\delta} dl^+_{\mu,\sigma}(\delta) = \frac{\mathrm{d}}{\mathrm{d}\delta} w_{i,j}(\delta - d_{i,j})^2$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\delta^2} dl^+_{\mu,\sigma}(\delta) = \frac{\mathrm{d}^2}{\mathrm{d}\delta^2} w_{i,j}(\delta - d_{i,j})^2$$



Optimizing PE using WM

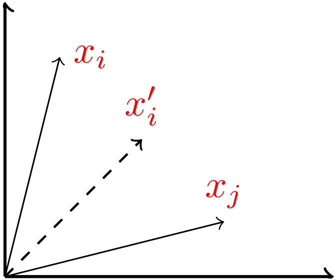
$$w_{i,j}^{+} = \frac{\frac{2}{\pi \ln 2} e^{-\frac{(\delta-\mu)^2}{\sigma^2}}}{\sigma^2 (1 - \mathrm{erf}(\frac{\delta-\mu}{\sqrt{2}\sigma}))^2} - \frac{\frac{1}{\sqrt{2\pi}\ln 2}(\delta-\mu) e^{-\frac{(\delta-\mu)^2}{2\sigma^2}}}{\sigma^3 (1 - \mathrm{erf}(\frac{\delta-\mu}{\sqrt{2}\sigma}))}$$

$$d_{i,j}^{+} = \delta - \frac{\frac{1}{\sqrt{2\pi}\ln 2} e^{-\frac{(\delta-\mu)^2}{2\sigma^2}}}{w_{i,j}\sigma (1 - \mathrm{erf}(\frac{\delta_{i,j}-\mu}{\sqrt{2}\sigma}))}$$

Update embeddings

$$x_i := \frac{\sum_{j \neq i} w_{i,j}(x_j + s_{i,j}(x_i - x_j))}{\sum_{j \neq i} w_{i,j}}, \quad (7)$$

where $s_{i,j} = \begin{cases} d_{i,j}/\|x_i - x_j\| & \text{if } \|x_i - x_j\| \neq 0 \\ 0 & \text{otherwise} \end{cases}.$



- $s_{i,j} = 0 \Rightarrow d_{i,j} = 0$
  $x_i \leftarrow x_j$
- $s_{i,j} = 1 \Rightarrow d_{i,j} = \|x_i - x_j\|$
  $x_i \leftarrow x_i$
- $s_{i,j} = \frac{1}{2} \Rightarrow d_{i,j} = \frac{1}{2}\|x_i - x_j\|$
  $x_i \leftarrow \frac{x_i + x_j}{2}$

Finding μ and σ

Meaning of μ: Boundary of Euclidean distance.

In the experiment, μ is fixed to a value between 1 and 2.

The optimal σ is found by a simple bisection search.

Algorithm

algorithm GEMPE $(V, E, d) \to \mathbb{R}^{n \times d}$

$\forall i \in V$: initialize $\mathbf{x}_i \in \mathbb{R}^d$ randomly (uniform distribution);

repeat

determine σ according to Section 3.2; → 每一轮中,找最优的 σ.

分子 $\forall i \in V$: $\mathbf{y}_i := (0, \ldots, 0)^\top \in \mathbb{R}^d$; // numerators and...

$\forall i \in V$: $z_i := 0 \in \mathbb{R}$; // ...denominators of new points

n是固定的

(*) **foreach** $(i, j) \in E$ **do**

分母

determine parabola $(d^+_{i,j}, w^+_{i,j})$ acc. to Eq. (3) and (4);

weightedMajorizationStep (i, j);

// negative sampling:

sample $k \in V$ such that $k \neq i$ and $(k, i) \notin E$;

determine parabola $(d^-_{k,i}, w^-_{k,i})$ acc. to Eq. (5) and (6);

i,j分别作 weightedMajorizationStep (k, i);

负采样

sample $k \in V$ such that $k \neq j$ and $(k, j) \notin E$;

determine parabola $(d^-_{k,j}, w^-_{k,j})$ acc. to Eq. (5) and (6);

weightedMajorizationStep (k, j);

$\forall i \in V$: $\mathbf{x}_i := \frac{1}{z_i} \cdot \mathbf{y}_i$; 更新.

until convergence;

return $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$;

Algorithm

procedure weightedMajorizationStep (a, b)
// one step of weighted majorization for $\mathbf{x}_a$ and $\mathbf{x}_b$:
$\mathbf{y}_a := \mathbf{y}_a + w_{a,b} \cdot \mathbf{x}_b;$
$\mathbf{y}_b := \mathbf{y}_b + w_{a,b} \cdot \mathbf{x}_a;$
$z_a := z_a + w_{a,b};$
$z_b := z_b + w_{a,b};$
if $||\mathbf{x}_a - \mathbf{x}_b|| \neq 0$ **then**
$\mathbf{y}_a := \mathbf{y}_a + w_{a,b} \cdot \frac{d_{a,b}}{||\mathbf{x}_a - \mathbf{x}_b||} \cdot (\mathbf{x}_a - \mathbf{x}_b);$
$\mathbf{y}_b := \mathbf{y}_b + w_{a,b} \cdot \frac{d_{a,b}}{||\mathbf{x}_a - \mathbf{x}_b||} \cdot (\mathbf{x}_b - \mathbf{x}_a);$

Experiments: Repr. Learning

Test embedding quality on downstream task: node classification.

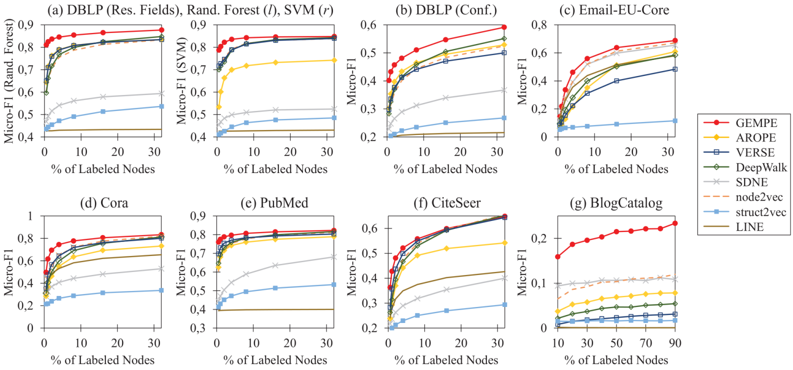


Figure 4: Comparison to Representation Learning Methods: Node Classification.

Experiments: Clustering

	DBLP *f.*	DBLP *c.*	Email	PubMed
GEMPE	**0.38**	**0.27**	0.67	0.29
AROPE	0.02	0.09	0.50	0.03
deepWalk	0.35	0.26	0.69	**0.30**
LINE	0.00	0.00	0.55	0.00
SDNE	0.02	0.04	0.62	0.02
node2vec	0.09	0.21	**0.70**	0.14
struct2vec	0.00	0.00	0.22	0.00
VERSE	0.36	0.25	0.67	0.28

Table 1: Comparison of Clustering Results (NMI) in 128D.

Experiments: Graph Drawing

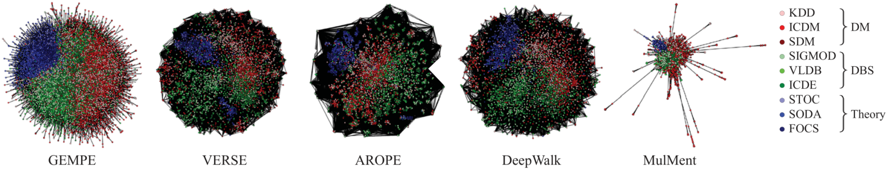


Figure 7: Visualization of DBLP Co-authorship Graph.

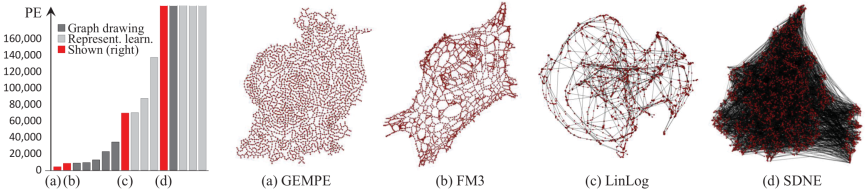


Figure 8: PE as a Quality Measure for Graph Drawing: Clear drawings of the Minnesota road network tend to have a low PE.

References I

Ulrik Brandes and Christian Pich. “An Experimental Study on Distance-Based Graph Drawing”. In: *Graph Drawing*. Ed. by Ioannis G. Tollis and Maurizio Patrignani. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 218–229. ISBN: 978-3-642-00219-9.

Emden R. Gansner, Yehuda Koren, and Stephen North. “Graph Drawing by Stress Majorization”. In: *Proceedings of the 12th International Conference on Graph Drawing*. GD’04. New York, NY: Springer-Verlag, 2004, pp. 239–250. ISBN: 3540245286. DOI: 10.1007/978-3-540-31843-9_25. URL: https://doi.org/10.1007/978-3-540-31843-9_25.