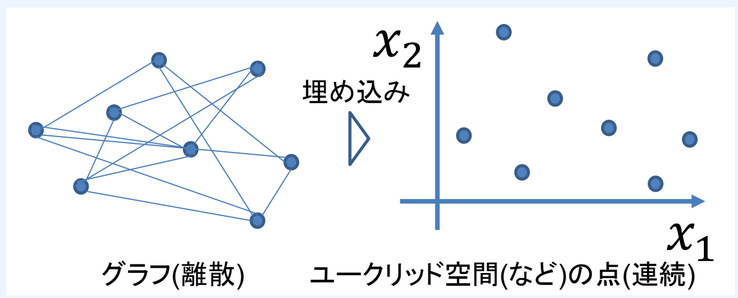
1-057: Dimensionality Selection of Hyperbolic Graph Embeddings using Decomposed Normalized Maximum Likelihood Code-Length

(MDL原理を用いた双曲埋め込みの次元選択; to be presented at ICDM2022)

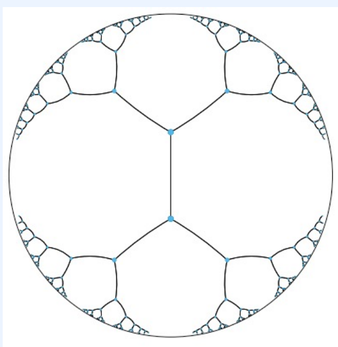
結城 凌 (東京大学); 池 祐一 (東京大学); 山西 健司 (東京大学)

研究目的

- グラフ埋め込み研究の発展 [1, 2].



- 埋め込み先の次元選択研究 [3, 4].
  - リンク予測・ノード分類の性能に影響.
  - ほとんどがユークリッド空間での研究.
- 双曲空間: 階層構造に適した空間*.



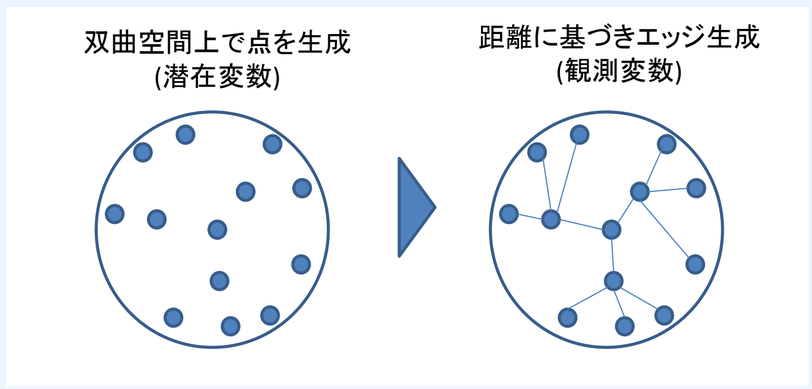
・左の線は双曲空間において等距離.

・表面積$\propto (\sinh r)^{D-1} = O(\exp r)$
・木のr番目のリーフ数$\propto 2^r$

本研究の貢献:双曲埋め込みでの次元選択手法の考案.

提案手法

- 潜在変数モデルを設定 [5].
  - ノード次数分布のべき則・高いクラスタ係数など.



- Minimum Description Length (MDL) 原理 [6]を用いる.
  - 統計的モデル選択基準の1つ. 与えられたデータを最も短く符号化するモデルを最良とする.
  - 一致性 [6]・PAC学習での早い収束 [7]など.
- Decomposed Normalized Maximum Likelihood (DNML)符号長 [8].
  - 潜在変数モデルで簡便に符号長を計算可能な手法.

$$L_{\mathrm{DNML}}(\boldsymbol{y}, \boldsymbol{z}) := L_{\mathrm{NML}}(\boldsymbol{y}|\boldsymbol{z}) + L_{\mathrm{NML}}(\boldsymbol{z}),$$
$$L_{\mathrm{NML}}(\boldsymbol{y}|\boldsymbol{z}) := -\log p(\boldsymbol{y}|\boldsymbol{z}; \hat{\beta}(\boldsymbol{y}, \boldsymbol{z})) + \log \sum_{\bar{\boldsymbol{y}}} p(\bar{\boldsymbol{y}}|\boldsymbol{z}; \hat{\beta}(\bar{\boldsymbol{y}}, \boldsymbol{z})),$$
$$L_{\mathrm{NML}}(\boldsymbol{z}) := -\log p(\boldsymbol{z}; \hat{\sigma}(\boldsymbol{z})) + \log \int p(\bar{\boldsymbol{z}}; \hat{\sigma}(\bar{\boldsymbol{z}})) d\bar{\boldsymbol{z}}_{1:D},$$

DNMLを最小化する次元Dを選択.

実験および結果

1. 人工データでの真の次元同定実験.

Benefit

# of nodes	DNML-HGG	AIC	BIC	MinGE
400	**0.042**	0.000	0.000	0.000
800	**0.250**	0.000	0.000	0.000
1600	**0.042**	0.000	0.000	0.000
3200	0.000	**0.250**	0.000	0.000
6400	**1.000**	0.375	0.000	0.000
12800	**1.000**	0.542	0.000	0.000

$$b(\hat{D}, D_{\mathrm{true}}) = \max\left\{0, 1 - \frac{|\log_2 \hat{D} - \log_2 D_{\mathrm{true}}|}{T_{\mathrm{gap}}}\right\}.$$

ノードが多い場合，提案手法は高い精度で真の次元に近い次元を推定.

2. 論文共著関係ネットワーク [9]でのリンク予測性能および推定次元の関係.

選択次元　最高性能

AUC

Network	2	4	8	16	32	64
AstroPh	0.549	0.813	0.842	0.846	0.849	**0.850**
CondMat	0.522	0.754	**0.767**	0.764	0.766	**0.767**
GrQc	0.526	0.728	0.744	**0.756**	0.755	**0.756**
HepPh	0.559	0.847	0.880	0.880	0.883	**0.885**

≥(最高性能)-0.01

提案手法は性能を保ちつつ最高性能を実現する次元に近い次元を推定.

3. WordNet [10]から生成したデータセットにおける，埋め込みの階層構造保存度および推定次元の関係.

Benefit

DNML-HGG	AIC	BIC	MinGE
0.714	0.286	0.500	0.000

提案手法は階層構造を最も保存する次元に近い次元を推定.

研究のポイント

- 双曲空間での次元選択手法構築.
  - 先行研究の少なさ [11].
- 潜在変数モデルおよびDNMLの利用.
  - 潜在変数モデルを用いない場合，Non-identifiability Problemにより符号長の解析計算が非常に困難.
- 人工データ・実データでの性能検証.

*画像は次のURLから引用: https://collignon.github.io/2020/07/notes-on-hyerpbolic-geometry/

[1] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, “Line: Large-scale information network embedding,” in Proceedings of the 24th international conference on world wide web, 2015, pp. 1067–1077.
[2] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–
864.
[3] Z. Yin and Y. Shen, “On the dimensionality of word embedding,” in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 895–906.
[4] P. T. Hung and K. Yamanishi, “Word2vec skip-gram dimensionality selection via sequential normalized maximum likelihood,” Entropy, vol. 23, no. 8, p. 997, 2021.
[5] W. Yang and D. Rideout, “High dimensional hyperbolic geometry of complex networks,” Mathematics, vol. 8, no. 11, p. 1861, 2020.
[6] J. Rissanen, Optimal estimation of parameters. Cambridge University Press, 2012.
[7] K. Yamanishi, “A learning criterion for stochastic rules,” Machine Learning, vol. 9, no. 2-3, pp. 165–203, 1992.
[8] K. Yamanishi, T. Wu, S. Sugawara, and M. Okada, “The decomposed normalized maximum likelihood code-length criterion for selecting hierarchical latent variable models,” Data Mining and Knowledge Discovery, vol. 33, no. 4, pp. 1017–1058, 2019.
[9] J. Leskovec and R. Sosi˘c, “Snap: A general-purpose network analysis and graph-mining library,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 8, no. 1, p. 1, 2016.
[10] G. A. Miller, WordNet: An electronic lexical database. MIT press, 1998.
[11] P. Almagro, M. Boguna, and M. Serrano, “Detecting the ultra low dimensionality of real networks,” arXiv preprint arXiv:2110.14507, 2021.