A-040

無順序木パターンクラスに対する変分グラフオートエンコーダの設計と構造解析
Design and Structural Analysis of a Variatioal Graph Autoencoder for Unordered Tree Pattern Classes

稗田 桃子 [1] 正代 隆義 [2] 松本 哲志 [3] 内田 智之 [4]
Momoko Hieda Takayoshi Shoudai Satoshi Matsumoto Tomoyuki Uchida

1. はじめに

深層生成モデルの発展に伴い，グラフ構造を対象とした表現学習およびグラフ生成モデルに関する研究が活発に進められている．その中でも Variational Graph Autoencoder (VGAE) [1]は，グラフの構造情報を効率的に潜在空間へ埋め込む手法として提案され，主に頂点表現の学習やリンク予測などに用いられてきた．一方，グラフ全体の生成に焦点を当てた GraphVAE [2]は，分子構造や構文木といった小規模なグラフの生成タスクにおいて，その有用性が広く認められている．

本研究は，こうしたグラフ生成モデルの研究動向を踏まえ，木構造の中でも特に無順序木(unordered tree)に注目する．無順序木は，頂点の子の順序を意味に含まない抽象的な構造であり，XML/HTML データ，論理式，構文パターン抽出などのさまざまな場面で自然に現れる．こうした構造を形式的に扱う枠組みとして，特定の無順序木パターンにマッチする構造の集合として定義される無順序木言語が以前から研究されてきた．我々の先行研究[3]では，こうした無順序木言語がグラフ畳み込みネットワーク(GCN)によって極めて高精度に学習可能であることを示しており，本研究はこの成果を踏まえ，無順序木言語に対する生成モデルとしての構成と評価を行う．

本研究では，線形無順序木パターンから定義される各無順序木言語を対象として，共通の構造を持つ変分オートエンコーダモデル LUTP-VAE を提案する．具体的には，各無順序木言語に対して同一構造のモデルを設定し，それぞれ個別に学習を行うことで，パターン構造に応じた潜在表現の獲得と妥当な無順序木の生成を目指す．モデル内部では，無順序木を完全グラフ構造に変換した特徴ベクトルを入力とし，潜在表現を経て出力された各頂点の特徴から親子関係を再構成する手法を用いており，出力構造の整合性も本モデルの設計において考慮されている．本論文では，その生成結果を元の言語に属するかどうかを判定するマッチ率や編集距離に基づいて評価する．

2. 無順序木パターン

本研究では，サイクルを持たない連結な無向グラフに対し，ある頂点を根として指定したものを**無順序木**と呼ぶ．無順序木では，同じ親を持つ子の間に順序関係はなく，親子関係のみが定義される．各頂点と各辺には，それぞれ有限集合Σ (頂点ラベル集合)およびΛ (辺ラベル集合)からの記号が付与される．また，Xを$\Sigma \cup \Lambda$と共通部分が空の無限集合とし，変数ラベル集合とする．ただし本研究では，$\Sigma = \{\varepsilon\}, \Lambda = \{\varepsilon\}$とする．ここで$\varepsilon$は空語(empty word)を表す．

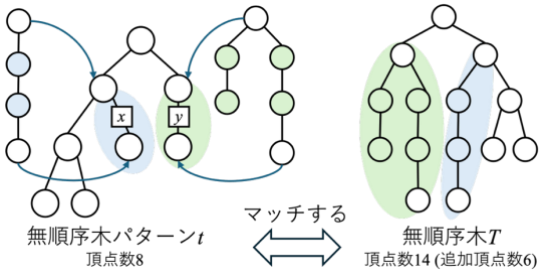

図 1: 無順序木パターン t がマッチする無順序木 T の例

無順序木パターンとは，無順序木の葉に接続する辺の一部を変数として指定し，それらを別の無順序木に置き換えることが可能な無順序木構造である．

束縛とは，変数に指定された辺を削除し，指定された無順序木の根をその辺の親頂点に，葉の一つを子頂点に対応させて貼り付ける操作をいう(詳細な定義は[3,4]を参照のこと)．束縛の集合を**代入**という．すべての変数が異なる変数ラベルを持つ無順序木パターンを**線形無順序木パターン**と呼び，本研究ではこの線形無順序木パターンのみを扱う．線形無順序木パターンの全体集合を$\mathcal{LUTP}_{\Sigma,\Lambda,X}$，無順序木の全体集合を$\mathcal{UT}_{\Sigma,\Lambda}$と表す．ここで，ラベル付き無順序木$T_1, T_2$が同型であるとは，両者の根が一致して，親子関係および各頂点・辺のラベルを保つようなT_1の頂点集合からT_2の頂点集合への全単射が存在することをいう．

代入θ (束縛の集合)によって得られる新たな無順序木を$t\theta$で表す．線形無順序木パターン$t \in \mathcal{LUTP}_{\Sigma,\Lambda,X}$に対して，すべての可能な代入$\theta$によって得られる無順序木の集合: $L(t) = \{t\theta \in \mathcal{UT}_{\Sigma,\Lambda} \mid \theta$は代入$\}$を，$t$によって定義される無順序木言語と呼ぶ．また，無順序木$T \in \mathcal{UT}_{\Sigma,\Lambda}$が，$T \in L(t)$であるとき，$T$は$t$に**マッチする**という(図 1)．以降，線形無順序木パターンを単に無順序木パターンと呼ぶ．

3. LUTP-VAE の構造

LUTP-VAE は，無順序木パターン$t \in \mathcal{LUTP}_{\Sigma,\Lambda,X}$に対応する無順序木言語$L(t)$の空でない有限部分集合$S \subseteq L(t)$を学習対象とし，各無順序木$T \in S$に対して再構成された無順序木を生成する深層生成モデルである．ここで，tの頂点数をp $(p \geq 1)$，Sに含まれる各無順序木の頂点数をn $(n \geq 1)$とする．すなわち，Sはすべて頂点数nの無順序木からなり，それぞれtの変数に代入された部分構造を含んでいる．

無順序木の入力表現とエンコーダ

無順序木Tは右深さ優先順序木表現に変換されたうえで，幅優先順に0から$n-1$まで頂点番号を付与する．ここで右深さ優先順序木表現とは，無順序木に対して深さ優先探索を行った際の各頂点の深さ列が辞書式順序で最小となるよ

1 福岡工業大学大学院工学研究科 Graduate School of Engineering, Fukuoka Institute of Technology
2 福岡工業大学情報工学部 Faculty of Information Engineering, Fukuoka Institute of Technology
3 東海大学理学部 Faculty of Science, Tokai University
4 広島市立大学大学院情報科学研究科 Graduate School of Information Sciences, Hiroshima City University

うな無順序木の順序木表現のことである．次に，同じ頂点数nを持つ完全グラフK_nを用意し，頂点番号を付与する．Tの頂点番号順に整列させた隣接行列Aの第i行$(1 \leq i \leq n)$を，K_n上の頂点番号$i-1$の特徴ベクトルとし，$n \times n$の入力特徴行列xを形成する．この完全グラフの構造と特徴行列が，LUTP-VAE のエンコーダへの入力となる．

エンコーダは，各頂点のn次元入力特徴ベクトルに対して，複数層の TAGConv [5]を適用し，最終的にd次元$(1 \leq d < n)$の潜在空間の平均ベクトル mean と対数分散ベクトル log_var を出力する．TAGConv は，グラフ上で各頂点の特徴を接続距離に応じて段階的に集約する畳み込み手法であり，本研究では集約深度として K=3 とした．ここで K は，各頂点が距離 0 から K までの近傍情報を取り込む最大伝播ステップ数である．

エンコーダが出力した mean と log_var から，再パラメータ化トリックにより潜在変数zをサンプリングする．標準正規分布に従うノイズを加えることで，連続的かつ微分可能な潜在表現を得る．

デコーダと無順序木構造の再構成

デコーダは，潜在変数zを入力として TAGConv 層を逆順に適用し，出力特徴行列$y \in [0,1]^{n \times n}$を得る．最終層ではシグモイド関数を適用し，出力を$[0,1]$の範囲に正規化する．

デコーダは，完全グラフの各頂点に対応するn次元の特徴ベクトルを出力する．各頂点$i > 0$に対して，ベクトル$y[i]$のうちインデックスがi未満の要素の中で最大の値を持つインデックスj^*を親と定める．この操作をすべての$i = 1, \ldots, n-1$に対して行うことで，根を0番とする木構造が一意に再構成される．子の順序に意味がないため，再構成されたグラフ構造は無順序木とみなすことができる．

LUTP-VAE は，再構成誤差と KL ダイバージェンスの加重和を最小化する目的関数により学習される．再構成誤差は，入力xと出力yの間のバイナリクロスエントロピーによって計算され，KL 項は，正規分布間の KL ダイバージェンスを解析的に導出した式に基づいて計算される．

4. 実験と評価

各実験において，無順序木パターン$t \in \mathcal{LUTP}_{\Sigma,A,X}$の頂点数$p$と，追加される構造部分の頂点数$n-p$を変化させた無順序木$T \in L(t)$を 100 パターンごとに 1000 個ずつ生成した．学習データには各パターンにつき 4000 個，評価には 1000 個の無順序木を用いた．潜在変数の次元は，入力特徴ベクトルの次元(=頂点数)に対して 1/4 とした．

まず，生成木が元の無順序木パターンtにマッチするかどうか($T' \in L(t)$であるか)を判定し，マッチ率を測定した．図 2 に，パターン頂点数$p \in \{5, \ldots, 15\}$，追加頂点数$n-p \in \{10, \ldots, 17\}$に対するマッチ率の平均を示す．マッチ率は全体として高く，特に$p \leq 8$において$n-p$に依らず 90%以上の値を維持している．$p \geq 9$では若干の低下が見られるが，構造の複雑性に対して高い精度を示している．

次に，パターン頂点数$p = 10$に固定し，追加頂点数$n-p \in \{10, \ldots, 17\}$に対する編集距離の評価を行った．ここでは，次の指標を用いて構造的な再構成の正確さを測定した:

- 平均編集距離率 (編集距離率 = 編集距離 / 頂点数),
- 平均同型率 (同型率 = 同型な無順序木の割合),
- 編集距離が 2 以下，3 以下，4 以下の割合の平均.

図 3 に結果を示す．平均編集距離率は常に 0.045 前後と非常に小さく，編集距離 2 以下の割合も 81%以上である．距離 4 以下に収まる割合は 98%以上を維持しており，提案モデルが良好な再構成性能を持つことが確認された．

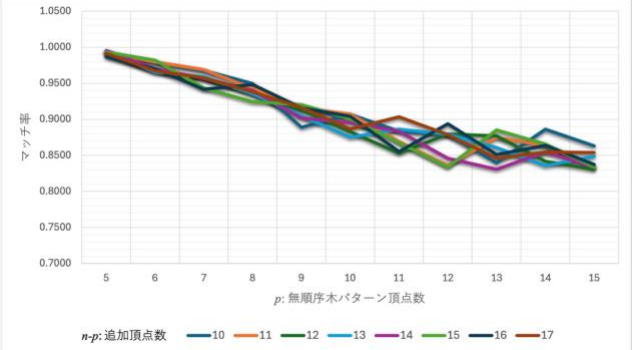

図 2: 再構成された無順序木のマッチ率による評価

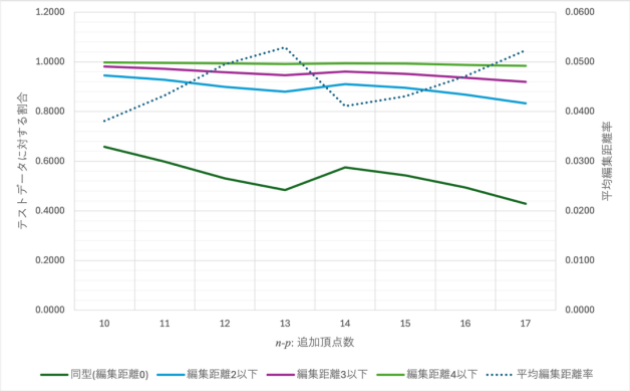

図 3: 再構成された無順序木の編集距離による評価($p = 10$)

5. おわりに

本研究では，無順序木パターン言語に基づく無順序木構造を対象とした深層生成モデル LUTP-VAE を提案した．完全グラフに基づく特徴表現と再構成手法により，マッチ率・同型率・編集距離において，再構成された無順序木の構造的類似性と近さを定量的に評価した．

また，入力無順序木の頂点数を次元とする入力特徴ベクトルに対し，その次元数に対する潜在変数の次元の割合を 1/6，1/4，1/2 とする複数構成において，モデルの比較も行っている．これらの設定による再構成性能，すなわち構造的類似性や近さへの影響については，発表時に詳細な評価を報告する予定である．

謝辞 本研究は JSPS 科研費 JP20K04973, JP21K12021, JP24K15074, JP24K15090 の助成を受けたものです．

参考文献

[1] Kipf, T. N. and Welling, M.: Variational Graph Auto-Encoders, arXiv:1611.07308 (2016).
[2] Simonovsky, M. and Komodakis, N.: GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders, Proc. ICANN 2018, LNCS vol.11139, Springer, pp.412-422, (2018).
[3] 石灘 洸樹，正代 隆義，内田 智之，松本 哲志：無順序木パターンに対する高精度グラフ畳み込みネットワークをオラクルとする質問学習モデルの解析，研究報告数理モデル化と問題解決 (MPS), 2023-MPS-143, 15, pp.1-8 (2023).
[4] Shoudai, T., Miyahara, T., Uchida, T., Matsumoto, S., and Suzuki, Y.: An Efficient Pattern Matching Algorithm for Unordered Term Tree Patterns of Bounded Dimension, IEICE Trans. Fundamentals, Vol.E101-A, No.9, pp.1344-1354 (2018).
[5] Du, J., Zhang, S., Wu, G., Moura, J. M. F., and Kar, S.: Topology Adaptive Graph Convolutional Networks, arXiv:1710.10370 (2017).