4M-06

超高精度グラフ畳み込みネットワークをオラクルとする無順序木パターンの質問学習モデル

石灘 洸樹[*1]　正代 隆義[*1]　内田 智之[*2]　松本 哲志[*3]
[*1]福岡工業大学情報工学部　[*2]広島市立大学大学院情報科学研究科　[*3]東海大学理学部

1. はじめに

木とはサイクルを持たない連結なグラフである．根付き木とは「根」と呼ばれる特別な頂点をただ一つだけ持つ木である．本論文では，順序木と呼ばれるデータ構造と区別するために，根付き木のことを無順序木と呼ぶ．小田ら[2]は，順序木データに対する二値分類問題を超高精度で分類するグラフ畳み込みネットワーク(GCN と略す)を報告した．本論文では，無順序木構造データに対する機械学習に関する問題を扱う．

質問学習モデルは Angluin[1]により提案された計算論的学習理論における機械学習モデルの一つである．質問学習モデルは，常に正答を返す教師(オラクルと呼ぶ)を仮定して計算量などの解析を行うモデルである．本論文では，GCN をオラクルとする無順序木パターンの質問学習モデルを提案する．そのモデル上で(1)無矛盾性問題，(2)二値分類問題，(3)可視化問題の 3 つの問題の精度を評価する．それにより，超高精度 GCN をオラクルとする質問学習手法の有効性を示す．

2. 無順序木パターン学習問題

2.1 無順序木パターンと代入操作

本論文ではグラフ理論の分野における用語を用いる．2 頂点以上の無順序木は根に基づく親子関係を持つ．根でない次数 1 の頂点を葉と呼ぶ．

頂点集合をV_T，辺集合をE_Tとする無順序木を$T=(V_T,E_T)$とする．Tの葉を端点とする辺の全体をE_T^{ℓ}とし，H_tをE_T^{ℓ}の部分集合とする．$V_t=V_T$と$E_t=E_T\setminus H_t$に対して，三つ組$t=(V_t,E_t,H_t)$を無順序木パターンと呼ぶ．H_tに属す辺を変数と呼ぶ．頂点と辺にはあらかじめ定められた有限アルファベットに属す記号(a,b,A,B など)が貼られているとする．また変数には互いに異なる変数ラベル(x,yなど)が貼られているとする．無順序木パターンは Shoudai ら[3]により提案された無順序項木パターンのサブクラスである．

変数ラベルxが貼られた無順序木パターンtの変数$h\in H$を無順序木Tで置き換える操作を，Tのxに対する束縛と呼ぶ．具体的には，xが貼られた変数hに対して，hの親である端点とTの根を同一視して，hを削除する．束縛の集合を代入と呼ぶ．tに代入θに属す束縛を全て実行して得られた無順序木パターンを$t\theta$と書く．無順序木パターンtと無順序木Tに対して，ある代入θが存在して，$t\theta$とTが根を根に対応させる写像により同型になるとき，tはTとマッチするという(図 1)．$L(t)$をtがマッチする全ての無順序木の集合とする．

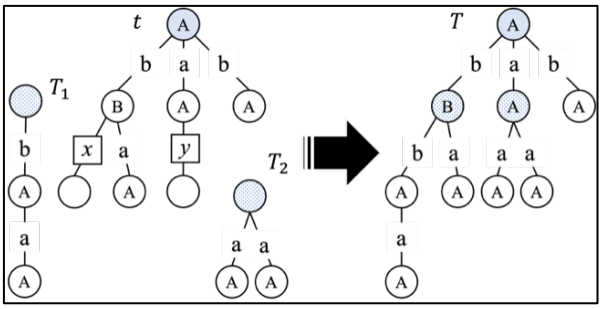

図 1: 無順序木パターンtとマッチするT

2.2 無順序木パターンに対する無矛盾性問題

S_+とS_-を無順序木の集合とする．ただし，ある無順序木パターンt_*が存在して，$S_+\subset L(t_*)$かつ$S_-\cap L(t_*)=\emptyset$であると仮定する．無順序木パターンに対する無矛盾性問題とは，仮定を満たすS_+とS_-を入力とし，$S_+\subset L(t)$かつ$S_-\cap L(t)=\emptyset$を満たす無順序木パターンtを出力する問題である．この問題は NP 完全である．

2.3 GCN における無順序木の二値分類

グラフ畳み込みネットワーク(GCN)の各層は，各頂点の特徴ベクトルを入力とし，隣接頂点の特徴ベクトルを反映した新たな特徴ベクトルを出力とする．最終的にそれらにより分類を行う．

無順序木パターンに対する無矛盾性問題の入力をS_+とS_-とする．GCN における無順序木の二値分類問題を，$S=S_+\cup S_-$を学習・検証・テストデータとして学習済み GCN (GCN^Sと記す)を構築する問題であるとする．S_+^GとS_-^GをGCN^SがSの無順序木をそれぞれS_+とS_-に属すと予測した無順序木の集合とする．GCN^Sの分類精度を，F 値(F_G)で評価する．$P_G=|S_+\cap S_+^G|/|S_+^G|$及び$R_G=|S_+\cap S_+^G|/|S_+|$とするとき，

$$F_G=2\cdot P_G\cdot R_G/(P_G+R_G).$$

A Query Learning Model for Unordered Tree Patterns with a Trained Ultra-High Precision GCN as an Oracle
*1 Hiroki Ishinada, Takayoshi Shoudai, Faculty of Information Engineering, Fukuoka Institute of Technology
*2 Tomoyuki Uchida, Graduate School of Information Sciences, Hiroshima City University
*3 Satoshi Matsumoto, Faculty of Science, Tokai University

GCN と質問学習の無順序木パターン発見手法

基本手続き

1. 無順序木パターンt_*を生成する．
2. ランダムにS_+とS_- ($S_+ \subset L(t_*)$かつ$S_- \cap L(t_*) = \emptyset$)を生成する．
3. $S = S_+ \cup S_-$を学習・検証・テストデータ(S_L, S_V, S_T)としてGCN^Sを構築する．F_Gを計算する．
4. GCN^Sをオラクルとし，全ての$T \in S_+$に質問学習を行う．$F_{Q(T)}$を計算する．
5. $F_{Q(T)}$の最大値F_Qを達成する無順序木パターンtを出力する．F_Qを計算する．

実験評価

6. ランダムに新データS'_+とS'_- ($S'_+ \subset L(t_*)$かつ$S'_- \cap L(t_*) = \emptyset$)を生成する．
7. $S' = S'_+ \cup S'_-$に対して，GCN^Sでの予測，無順序木パターンtでの分類を行い，F 値(F'_G, F'_{GQ}, F'_Q)を計算する．

図 2: GCN と質問学習による無順序木の学習

2.4 GCN と質問学習による無矛盾性問題の解法

無順序木パターンt_*に対する所属性質問とは，無順序木Tを入力とし，$T \in L(t_*)$か否かを答える質問である．任意の無順序木$T \in L(t_*)$を入力とし，何回かTを変換し，それに対して所属性質問を用いて，無順序木パターンt_*を同定する．

GCN^Sによるオラクルでは無順序木パターンt_*の同定に至らない可能性があることに注意する．無順序木$T \in S_+$に対してGCN^Sによる質問学習が出力する無順序木パターンをt_Tとする．$S_+^{Q(T)} = S \cap L(t_T)$, $S_-^{Q(T)} = S \setminus L(t_T)$とする．$t_T$の精度を F 値$(F_{Q(T)})$で評価する．$P_{Q(T)} = |S_+ \cap S_+^{Q(T)}|/|S_+^{Q(T)}|$及び$R_{Q(T)} = |S_+ \cap S_+^{Q(T)}|/|S_+|$とするとき，

$$F_{Q(T)} = 2 \cdot P_{Q(T)} \cdot R_{Q(T)}/(P_{Q(T)} + R_{Q(T)}).$$

最後に$F_{Q(T)}$の最大値$F_Q = \max_{T \in S_+} F_{Q(T)}$を達成する無順序木パターン$t$を出力する．

2.5 GCN の質問学習モデルによる可視化

小田ら[2]は質問学習モデルを GCN の判断根拠を可視化するための順序木構造パターンを出力する可視化手法であるとみなした．本論文では，GCN^Sの可視化としての本手法の精度を，新たに生成した無順序木データ$S' = S'_+ \cup S'_-$を用いて，F 値(F'_{GQ})で評価する．$P'_{GQ} = |S'^G_+ \cap S'^Q_+|/|S'^Q_+|$及び$R'_{GQ} = |S'^G_+ \cap S'^Q_+|/|S'^G_+|$とするとき，

$$F'_{GQ} = 2 \cdot P'_{GQ} \cdot R'_{GQ}/(P'_{GQ} + R'_{GQ}).$$

3. 学習実験と考察

本論文で提案する GCN と質問学習による無順序木の学習手法を図 2 にあげる．計算機実験では，

$$|S_+| = 5{,}000, |S_-| = 5{,}000, |S_L| = 6{,}400,$$
$$|S_V| = 1{,}600, |S_T| = 2{,}000, |S'| = 10{,}000$$

とし，この設定での実験を 1,500 回繰り返した．F 値$(F_G, F_Q, F'_G, F'_{GQ}, F'_Q)$ を横軸に，その F 値以上を達成した実験の割合を縦軸にしたグラフを図 3 に示す．1,500 回の実験のうち，F 値の最大値は，$F_G, F_Q, F'_G, F'_{GQ}, F'_Q$のいずれも 1.00 であった．平均値は$F_G = 0.9998, F_Q = 0.8030, F'_G = 0.9147, F'_{GQ} = 0.7253, F'_Q = 0.7949$であった．$GCN^S$は 66.533%の実験で$F_G = 1.00$を達成し，全実験で$F_G \geq 0.979$であり，$GCN^S$の分類精度は高い．一方，新データ$S'$に対して$F'_G = 1.00$を達成した実験の割合は低い．よって過学習が起こっていることが観測される．無順序木パターンtでの分類は，GCN^Sをオラクルとしたにも関わらず，新データでも元データとほぼ同様 F 値に対する実験数の割合が変化しており過学習は観測されない．これは，無順序木パターンであることを前提として質問学習を行っているからであると考えられる．

今後の課題は，無順序木のグラフ構造に注目して学習精度と計算量の解析を行うことである．

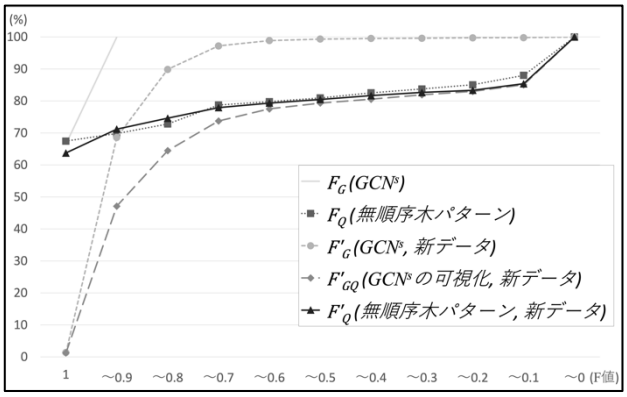

図 3: 横軸の F 値以上を達成した実験の割合

謝辞 本研究は JSPS 科研費 19K12103, 21K12021 の助成を受けたものです．

参考文献

[1] D. Angluin, Queries and Concept Learning, *Machine Learning*, 2(4), 319-342, 1988.
[2] 小田 直季 他, 順序木パターンの質問学習アルゴリズムによるグラフ畳み込みネットワークの予測根拠の可視化, 2022 年度 人工知能学会全国大会(第 36 回), 2G4-GS-2-01, 2022.
[3] T. Shoudai et al., An Efficient Pattern Matching Algorithm for Unordered Term Tree Patterns of Bounded Dimension, *IEICE Trans. Fundamentals*, E101.A(9), 1344-1354, 2018.