4M-05

学習済超高精度 GCN をオラクルとする順序項木パターンの質問学習モデルの解析と実データでの評価

東山 的生[*1]　野口 大悟[*2]　内田 智之[*3]　正代 隆義[*2]　松本 哲志[*4]

[*1]広島市立大学情報科学部　[*2]福岡工業大学情報工学部　[*3]広島市立大学大学院情報科学研究科　[*4]東海大学理学部

1. はじめに

深層学習は，高い予測能力を有しているため多くの分野で活用されている．その一方で，予測根拠に関する透明性・説明性が低い学習モデルであることも知られている．

Angluin[1]により提唱された質問学習モデルは，学習者が常に正答を返す教師(オラクル)に質問を繰り返すことで，教師の有する概念を同定する学習手法である．小田ら[2]は，順序木データを学習した高精度なグラフ畳み込みネットワーク(GCN)モデルをオラクルとし，順序木データに共通する概念の表現である順序項木パターンを獲得する質問学習アルゴリズムを提案した．

本稿では，小田ら[2]の質問学習アルゴリズムにより獲得した順序項木パターンについて二値分類精度である F 値に関して分析することで，超高精度 GCN モデルをオラクルとした質問学習アルゴリズムを評価する．さらに，Wikipedia などの Web ページを学習させた超高精度 GCN モデルをオラクルとした質問学習アルゴリズムにより獲得した高い F 値を有する順序項木パターンについて報告する．これにより，実データに対する超高精度 GCN をオラクルとした質問学習アルゴリズムの有効性を示す．

2. 準備

2.1 順序項木パターン

Σを有限アルファベット，$\Sigma \cap X = \emptyset$を満たす$X$を無限アルファベットとする．根をもち，すべての子が順序づけられている木を順序木という．V_tを頂点集合とし，$E_t \subseteq V_t \times (\Sigma \cup X) \times V_t$を辺集合とする順序木$t = (V_t, E_t)$を順序項木パターンという．辺$e = (u, a, v) \in E_t$における$a \in \Sigma \cup X$を辺$e$の辺ラベルという．特に$a \in X$である辺ラベル$a$を変数ラベルといい，変数ラベルを持つ辺を変数という．各変数ラベルが高々1 回しか現れない順序項木パターンを線形順序項木パターン(あるいは，順序項木パターン)といい，変数を持たない順序項木パターンを単に順序木ということにする．例として，図 1 に順序項木パターンtと順序木T, T'を示す．図において，項木パターンの変数は変数ラベルを四角で囲って表している．順序項木パターン全体の集合をOTPで，順序木全体の集合をOTで表す．順序項木パターン$t \in OTP$と順序木$T \in OT$に対して，tの各変数をそれぞれ適当な順序木で置き換えて得られる順序木がTと同型なとき，tとTはマッチするという．例えば，図 1 の順序項木パターンtの頂点v_1，v_2と順序木T'の頂点u_1，u_2を同一視してtの変数(v_1, x, v_2)を置き換えることで順序木Tと同型な順序木が得られるため，tとTはマッチする．

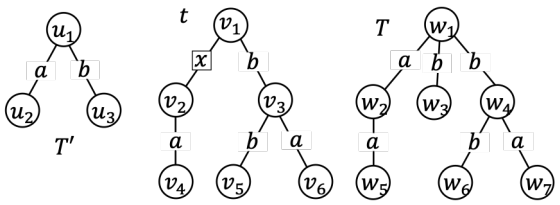

図 1: 順序項木パターンtと順序木T, T'

2.2 順序木集合に対する高精度 GCN モデル

GCN(Graph Convolutional Network)は，グラフデータに対して畳み込みを行う手法である．畳み込みとは，目標頂点の特徴ベクトルの更新の際に，隣接頂点の情報を取り込む手法である．つまり，各層において入力となる各頂点の特徴ベクトルに隣接頂点の特徴量を反映した新たな特徴ベクトルを割り当て，GCN の多層を経て最終的に得られた特徴ベクトルを用いて分類を行う手法である．

本稿では，二値分類された順序木集合を学習させた高精度 GCN モデル，RGCN(Relational Graph Convolutional Network) と GConvLSTM(Graph Convolutional Long Short-Term Memory)をオラクルとして用いる． RGCN は，頂点と隣接頂点間の

Analysis of a Query Learning Model for Ordered Term Tree Patterns with a Trained Ultra-High Precision GCN as an Oracle and Its Evaluation on Real Data

*1 Matoi Higashiyama, Faculty of Information Sciences, Hiroshima City University
*2 Daigo Noguchi and Takayoshi Shoudai, Faculty of Information Engineering, Fukuoka Institute of Technology
*3 Tomoyuki Uchida, Graduate School of Information Sciences, Hiroshima City University
*4 Satoshi Matsumoto, Faculty of Science, Tokai University

辺の種類を考慮した上で特徴ベクトルを集約することができるように GCN を拡張したモデルである．GConvLSTM は，長期的な依存関係を学習できる LSTM を GCN に取り入れたモデルである．

2.3 GCN モデルをオラクルとする質問学習手法

小田ら[2]は，二値分類された順序木集合Dを学習させた高精度 GCN モデルMをオラクルとして用い，D中の順序木に対するMによる分類予測の根拠を表す順序項木パターンを出力する質問学習アルゴリズムを提案した．このアルゴリズムでは，Dの一方の分類に入る順序木を正例，もう一方の分類に入る順序木を負例とし，D中の各順序木に対して(1) 最小正例発見処理と(2)変数同定処理を施す．(1)では，順序木$t \in D$の各辺に対して縮約操作を行なって得られた順序木が正例となるかをオラクルMに問い合わせる．”yes”が返ってきたら縮約して得られた順序木に，”no”が返ってきたら縮約前の順序木に再帰的に縮約操作を適用して最小辺数の正例である順序木を求める．さらに，(2)では，最小変数の正例の各辺を小さな負例で置き換えた順序木が正例か否かをオラクルMに問い合わせ，”yes”ならば変数とし，”no”ならば辺に戻す操作を再帰的に行って順序項木パターンを生成する．

3. 実験と考察

二値分類された順序木集合S_+とS_を学習データとする GCN モデルMおよび質問学習により得られた順序項木パターンtの二値分類精度である F 値 F_MおよびF_tを次のように定義する．$S_+ \cup S_-$中の順序木のうち，MがそれぞれS_+とS_-に属すると予想したものの集合をそれぞれS_+^GとS_^Gで表すとき，$F_M = 2 \cdot P_M \cdot R_M/(P_M + R_M)$と定義する．ここで，$P_M = |S_+ \cap S_+^G|/|S_+^G|$, $R_M = |S \cap S_+^G|/|S_+|$である．同様に，$S_+ \cup S_-$中の順序木のうち$t$とマッチするものとマッチしないもの集合をそれぞれ$S_+^G$と$S_-^G$としたときの$F_M$の値を$F_t$とする．

本稿では，次に示す(1) 二値分類された順序木集合(人工データ)と (2)インターネット上のWebページの集合（実データ）をそれぞれ学習させた高精度 RGCN モデル $RGCN_1$ と $RGCN_2$および高精度 GConvLSTM モデル $GConvLSTM_1$ と $GConvLSTM_2$ をそれぞれ構築し質問学習の実験を行った．(1)では，図 2 の順序項木パターンtとマッチする 4000 個の順序木の集合S_+とマッチしない 4000 個の順序木の集合S_-を人工的に作成し，$S_+ \cup S_-$を 6400 個の訓練データと 1600 個のテストデータに分けて学習させた，F 値がともに 1.0 の $RGCN_1$ と $GConvLSTM_1$ をそれぞれ構築した．$RGCN_1$ と $GConvLSTM_1$ をオラクルとした質問学習アルゴリズムにより得られた順序項木パターンのうち F 値

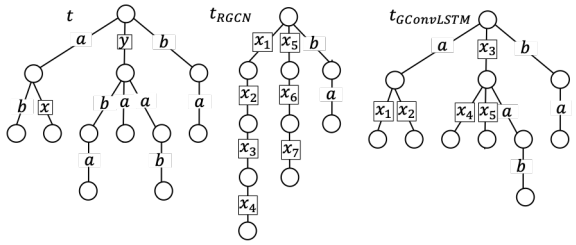

図 2：順序項木パターン$t, t_{RGCN}, t_{GConvLSTM}$

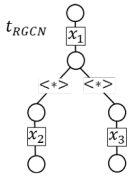

図 3：順序項木パターン t_{RGCN}

0.98 のt_{RGCN}と F 値 1.0 の$t_{GConvLSTM}$をそれぞれ図 2 に示す．t_{RGCN}と$t_{GConvLSTM}$の両方とも F 値は高いが，$t_{GConvLSTM}$の方が順序項木パターンtの構造をより詳細に掴んでいることが図 2 から見て取れる．

(2) では，Web スクレイピングの手法を用い，良質な記事が掲載された Wikipedia 内の記事リンクから得られる 1806 個のページの集合S_+と日本株の配当金データベースから得られる上場企業のホームページ 1000 個の集合S_-を 1796 個の訓練データ，449 個の検証データ，561 個のテストデータの学習データとして F 値がともに 1.0 の $RGCN_2$ と $GConvLSTM_2$ を構築した．この $RGCN_2$ をオラクルとした質問学習アルゴリズムにより得られた F 値 0.97 の順序木パターンを図 3 に示す．この結果は，実データに対しても高精度 GCN モデルをオラクルとする質問学習アルゴリズムが有効であることを示唆している．

4. おわりに

本稿では，順序木集合(人工データ)と Web ページ(実データ)それぞれを学習させた高精度 GCN をオラクルとした質問学習手法について，獲得した順序項木パターンの F 値と構造について分析することで，その質問学習手法の有効性を示した．本研究は JSPS 科研費 19K12103，21K12021 の助成を受けたものである．

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.