様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 １２ 日現在

機関番号：１２６０１
研究種目：研究活動スタート支援
研究期間：2021～2022
課題番号：２１Ｋ２１２８０

研究課題名（和文）大規模不均衡データ学習に対する新たなグラフニューラルネットワークの研究

研究課題名（英文）Graph Neural Networks for Large-Scale Imbalanced Data

研究代表者
鈴村　豊太郎（Suzumura, Toyotaro）

東京大学・情報基盤センター・教授

研究者番号：７０５５２４３８

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：本研究では、実世界アプリケーションのデータ表現として用いられるグラフ構造の中でも、不均衡ラベルを持つグラフデータ、長時系列の動的グラフデータの特徴を捉えるGNNモデルとソフトウェアアーキテクチャの最適化を提案した。不均衡ラベルに対しては、ヘテロジニアスグラフを構築し、ヘテロジニアスGNNモデルで学習する方法を提案することで、金融取引ネットワークの不正アカウント予測において高い性能を実現した。また、動的な大規模グラフの特徴を長期的なコンテキストで捉えるSpectral Waveletを、トポロジカルデータ解析で知識グラフの関係性の補完を効率的に評価する手法をそれぞれ提案した。

研究成果の学術的意義や社会的意義
大規模グラフデータを扱う機械学習などの研究では、動的なグラフ構造の変化、インバランスなラベルが要因となるモデル性能の問題を解決することは必須の課題である。本研究と並行して企業（自動車会社、新聞社、人材紹介会社）との共同研究を進めていく中でもこれらの課題が本質的な課題であることを確認しており、学術的な意義ばかりではなく社会的にも大きな意義のある研究成果と言える。

研究成果の概要（英文）：In this study, we proposed the optimization of a GNN (Graph Neural Network) model and software architecture to capture the characteristics of graph data with imbalanced labels and long-term dynamic graph data, which are used as data representations in real-world applications. For imbalanced labels, we proposed a method to construct a heterogeneous graph and train it with a heterogeneous GNN model, which achieved high performance in predicting fraudulent accounts in financial transaction networks. Additionally, we proposed the Spectral Wavelet, which captures the characteristics of dynamic large-scale graphs in a long-term context, and an efficient method for evaluating the complementarity of knowledge graph relationships using topological data analysis.

研究分野：人工知能

キーワード：グラフニューラルネットワーク　機械学習　人工知能　グラフ解析　大規模データ　高性能計算

１．研究開始当初の背景

ノード（頂点）と、ノード同士を接続するエッジ（枝）から構成されるグラフ構造は、現実世界、デジタル空間、生体内、自然界におけるデータをより直感的に表現することができ、幅広く用いられるデータ構造である。その中でも、ノード数が数百万から数十億に至るまで大規模であり、ノードやエッジ自身がラベルなどの属性を持っていたり、構造が時系列および空間的な意味を持ったりする。これらのグラフ構造に対して、グラフ内のノードやエッジ、部分グラフ、もしくはグラフ全体の属性を機械学習のラベルとみなして分類する手法の研究が盛んに行われている。特にグラフニューラルネットワーク（Graph Neural Network）GNN は、ノードが持つ特徴量を隣接ノードに伝播させることで、ニューラルネットワークを用いて自動的にグラフ構造を学習するモデルであり、教師付き機械学習、クラスタリングや異常検知などの教師なし学習に繋げる手法として、最も注目を浴びている領域である。

一方、予測すべきラベルの分布が著しく不均衡なデータセットに対して、ラベル予測などのタスクを行う不均衡データ学習も行われているが、不正などを示すラベルが全体の中で極端に少ないため従来の機械学習および GNN では十分な精度と実行性能が得られない問題がある。代表的なアプリケーションとして、金融決済データにおける不正パターンの検出が挙げられ、人間が不正ラベルと判定したものを用いることで、教師付き機械学習により不正なノードやエッジの特性を判定することができる。しかし、不正取引は全体の中でもごく少数であり、高い精度で検出するのは困難である。このような背景の中で、不均衡データ学習に対する GNN モデルのアーキテクチャと、それを包含する汎用ソフトウェアアーキテクチャとはどうあるべきかが問われている。

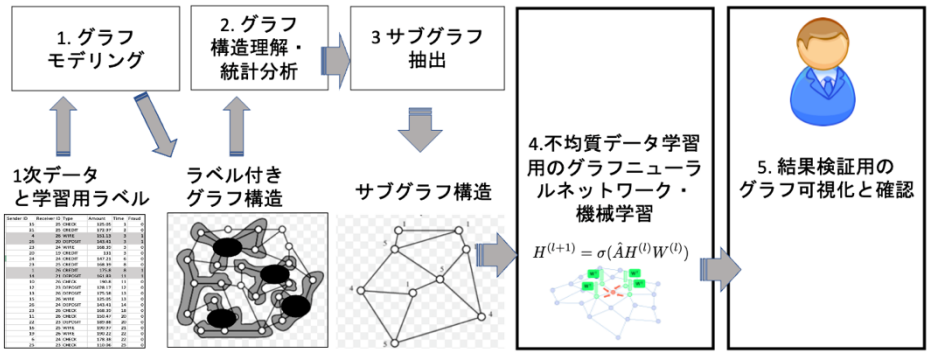
２．研究の目的

本研究の目的は、不均衡データ学習に対する新たな GNN 及びそれを包含するソフトウェアアーキテクチャを提案することである。アーキテクチャには GNN モデルに加え、グラフモデリング、前処理、後処理、人間によるラベリングの効率化等を含めた一連のフローがあるが、様々な応用問題に対して一般化できるアーキテクチャはまだ確立されていない。

GNN モデルに関しては、GraphSAGE, RGCN など様々なモデルが数多く提案されているが、不均衡データ学習に特化したような GNN モデルはまだ提案されていない。また、不均衡データ学習問題に対して、ソフトウェア工学的に上記のような機械学習の一連の工程をどのようなアーキテクチャで実現すべきか未解決である。そこで、不均衡グラフデータの特性を把握し、最適な GNN モデルおよびソフトウェアアーキテクチャを設計し提案することを目指す。

３．研究の方法

本研究では、様々な実問題とデータセットに対応する、不均衡データ学習のための GNN とそれを支えるソフトウェアアーキテクチャがどうあるべきかを明らかにする。具体的には、下図のようなアーキテクチャの骨子を元にして、以下に述べるそれぞれの Phase について実装・評価を繰り返しながら検証を行い、その有効性を明らかにしていく。




Phase　1:　適切なグラフモデリングの設計　最初に、実問題より取得した１次データをグラフ構造に変換するグラフモデリングを行う。ここでは一次データのどの要素をノード、エッジ、機械学習用の学習ラベル及びノードの属性として使用するかを決める。

Phase　2:　グラフとラベル付きノードの理解　構築されたグラフに対して効果的にラベル付きノードの学習を行えるよう、グラフ中のラベル付きノードの特徴について解析する。例えば、グラフを複数の連結成分に分割した上で、連結成分ごとのラベル付きノードの分布などの統計値を集計することで、機械学習において効果が期待できる特徴量を選ぶための指針にする。

Phase　3.　サブグラフの抽出　最もラベル付きノードが存在する連結成分を抽出し、ラベル付きノードの周辺からなる部分グラフのみを学習対象とする。

Phase 4: 大規模不均衡グラフ用 GNN による学習 ラベルの不均衡問題を軽減しつつ、大規模なグラフデータを効率的に処理し精度を向上させるため、Phase 2 の解析結果を踏まえた GNN を提案し、ラベルの予測を行う。Phase 5 で人間が結果を検証することを踏まえ、予測結果のラベルだけではなく該当ノードの特徴量、ラベル付きノード周辺の情報も含めて出す。
Phase 5: 最終判定を行う人間とのインタフェース ラベル付きと予測されたノードを参照しながら、周辺のノードの属性及びその他統計的な情報によって、予測結果の妥当性を判定する。

## ４．研究成果

不均衡なグラフデータに GNN モデルをそのまま適用した場合、ほとんどの正常なラベルを持つノードの特徴が多く伝わり、不正アカウントのような特殊なラベルを持つノードの特徴が隠れてしまい、これらのラベルの予測が難しくなる。そこで、ノードの役割に関する属性を特徴ではなくノードの「タイプ」として定義し、エッジも両端のノードのタイプによって定義されるヘテロジニアスグラフを構築した。さらに、異なるノードとエッジのタイプを区別するヘテロジニアス GNN モデルを用いて、各ノードの重要な情報を保ちながら GNN モデルを学習させる方法を提案した。金融ネットワーク上の不正アカウントを予測する課題に対して、アカウントの種類をノードタイプとしたヘテロジニアスグラフとヘテロジニアス GNN モデルを組み合わせた結果、ヘテロジニアス GNN モデル（右表 HAN, HGT, RGCN）はすべてのノードを同じタイプとする GNN モデル（右表 GCN, GAT, SAGE）と比べて高いモデル性能を実現し、特にエッジタイプごとに GCN モデルを学習する RGCN モデルが最高のモデル性能を達成した。この成果は、国際学会 KDD のワークショップ MLG で発表された。

| Model | Precision | Recall | F1 | PR-AUC |
|---|---|---|---|---|
| GCN | 0.8817 | 0.4387 | 0.5843 | 0.7154 |
| GAT | 0.8695 | 0.5093 | 0.6399 | 0.7558 |
| SAGE | 0.8244 | 0.6644 | 0.7354 | 0.8150 |
| HAN | 0.8800 | 0.4259 | 0.5725 | 0.6561 |
| HGT | 0.7718 | **0.7940** | 0.7849 | 0.8383 |
| RGCN | **0.8958** | 0.7454 | **0.8124** | **0.8896** |

研究としては上記の不均衡問題の他に、時系列・動的に変化する大規模グラフに対する GNN モデルの研究へと発展した。動的グラフに対応する GNN モデルはすでに数多く提案されているが、いずれも短期的なデータの変化しか考慮されておらず、実世界で扱われている長期的なグラフデータでは長期的なコンテキストを捉えることができない問題が潜在的に存在していた。この問題に対して、時間幅が非常に長いグラフデータの性質も捉えることができる Spectral Wavelet を提案し、国際学会 AAAI に採択された。また、知識グラフ上で足りない関係性を補完する手法を評価する方法として トポロジカルデータ解析（Topological Data Analysis）における Persistent Homology の概念を用いて効率的に評価する手法を提唱した。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　1件）

| | |
|---|---|
| １．著者名<br>Bastos, Anson, Abhishek Nadgeri, Kuldeep Singh, Hiroki Kanezashi, Toyotaro Suzumura, and Isaiah Onando Mulang | ４．巻<br>- |
| ２．論文標題<br>How Expressive are Transformers in Spectral Domain for Graphs | ５．発行年<br>2022年 |
| ３．雑誌名<br>Transactions on Machine Learning Research (2023) | ６．最初と最後の頁<br>1-8 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>無 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　3件）

| |
|---|
| １．発表者名<br>Anson Bastos, Kuldeep Singh, Abhishek Nadgeri, Johannes Hoffart, Manish Singh, and Toyotaro Suzumura |
| ２．発表標題<br>Can Persistent Homology provide an efficient alternative for Evaluation of Knowledge Graph Completion Methods? |
| ３．学会等名<br>ACM Web Conference 2023 (WWW '23)（国際学会） |
| ４．発表年<br>2023年 |

| |
|---|
| １．発表者名<br>Bastos, Anson, Abhishek Nadgeri, Kuldeep Singh, Toyotaro Suzumura, and Manish Singh |
| ２．発表標題<br>Learnable Spectral Wavelets on Dynamic Graphs to Capture Global Interaction |
| ３．学会等名<br>The 36th AAAI Conference on Artificial Intelligence (AAAI 2023)（国際学会） |
| ４．発表年<br>2023年 |

| |
|---|
| １．発表者名<br>Kanezashi, Hiroki, Toyotaro Suzumura, Xin Liu, and Takahiro Hirofuchi. |
| ２．発表標題<br>Ethereum Fraud Detection with Heterogeneous Graph Neural Networks. |
| ３．学会等名<br>28TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ' 22), Workshop on Mining and Learning with Graph (国際学会) |
| ４．発表年<br>2022年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|
| 研究協力者 | 金刺　宏樹<br>(Kanezashi Hiroki)<br>(80889395) | 東京大学・情報基盤センター・特任研究員<br>(12601) | |

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 | | | |
|---|---|---|---|---|
| インド | Indian Institute of Technology | | | |
| ドイツ | RWTH Aachen, Germany | | | |