論 文 内 容 の 要 旨

氏 名 (前川 政司)	
論文題名	Towards Practical Node Classification for Attributed Graphs: Improving Effectiveness/Scalability and Benchmarking Graph Neural Network-based Methods (実践的な属性付きグラフにおけるノード分類に向けた精度とスケーラビリティの向上およびグラフ深層学習手法の実証研究)

論文内容の要旨

Graphs appear everywhere in many application domains such as web page links, social networks, computer vision, and gene expressions. Graph processing attracts broad attention and node classification is one of the hottest topics in the graph machine learning field. Machine Learning (ML) methods including Graph Neural Networks (GNNs) are powerful tools for node classification. However, to apply such ML methods to practical applications, several issues remain: hardly leveraging class structure, limited scalability, no comprehensive evaluation, limited data types, limited support for time-series of graphs, and no general-purpose pre-trained models.
The first three limitations regarding effectiveness, scalability, and evaluation are fundamental to all the limitations since basic techniques and frameworks which address the three limitations can be extended to more complicated and practical settings, i.e., graphs with multiple node/edge types and/or time-series information. As for general-purpose pre-trained models, effectiveness, scalability, and evaluation are essential to ensure the model capability, adequate learning within a reasonable time, and the generalizability of the model, respectively. In this sense, overcoming the effectiveness, scalability, and evaluation limitations will be beneficial in overcoming the other limitations.
Hence, we addressed these three limitations in this thesis.
This thesis consists of five chapters. First, we describe the research background and discuss prior research works and their limitations in Chapter 1. In Chapter 2, we consider the clustering problem of attributed graphs in which we need to leverage the class structure, i.e., the relationship between the classes, attributes, and topology, in order to achieve high-quality performance. Note that graph clustering can be regarded as unsupervised node classification by assuming that nodes with the same label form a community. Our challenge is how we design an effective clustering method that captures the complicated relationship between the topology and the attributes in real-world graphs. We propose NAGC, a new attributed graph clustering method that bridges the attribute space and the topology space. The feature of NAGC is two-fold; 1) NAGC learns a projection function between the topology space and the attribute space so as to capture their complicated relationship, and 2) NAGC leverages the positive unlabeled learning to take the effect of partially observed positive edges into the cluster assignment. We conducted experiments extensively to validate that NAGC performs higher than or comparable to prior arts regarding the clustering quality.
In Chapter 3, we propose a framework that automatically transforms non-scalable GNNs into precomputation-based GNNs which are efficient and scalable for large-scale graphs. The advantages of our framework are two-fold; 1) it transforms various non-scalable GNNs to scalable ones so that the transformed ones scale well to large-scale graphs by separating local feature aggregation from weight learning in their graph convolution, 2) it efficiently executes precomputation on GPU for large-scale graphs by decomposing their edges into small disjoint and balanced sets. Through extensive experiments with large-scale graphs, we demonstrate that the transformed GNNs run faster in training time than the original GNNs while achieving competitive accuracy to the state-of-the-art GNNs. Consequently, our transformation framework provides simple and efficient baselines for future research on scalable GNNs.
In Chapter 4, we propose an evaluation framework using synthetic graphs for graph machine learning methods. First, we propose GenCAT, an attributed graph generator for controlling those relationships, which has the following advantages; 1) GenCAT generates graphs with user-specified node degrees and flexibly controls the relationship between nodes and labels by incorporating the connection proportion for each node to classes. 2) Generated attribute values follow user-specified distributions, and users can flexibly control the correlation between the attributes and labels. 3) Graph generation scales linearly to the number of edges. GenCAT is the first generator to support all three of these practical features. Through extensive experiments, we demonstrate that GenCAT can

efficiently generate high-quality complex attributed graphs with user-controlled relationships between labels, attributes, and topology. Second, we conduct extensive experiments with a synthetic graph generator that can generate graphs having controlled characteristics for fine-grained analysis. Our empirical studies clarify the strengths and weaknesses of GNNs from four major characteristics of real-world graphs with the class labels of nodes, i.e., 1) class size distributions (balanced vs. imbalanced), 2) edge connection proportions between classes (homophilic vs. heterophilic), 3) attribute values (biased vs. random), and 4) graph sizes (small vs. large). In addition, to foster future research on GNNs, we publicly release our codebase that allows users to evaluate various GNNs with various graphs. We hope this work offers interesting insights for future research.

Finally, Chapter 5 summarizes this thesis and discusses our future work.

様式 7

論文審査の結果の要旨及び担当者

氏　名（　前川　政司　）			
論文審査担当者		（職）	氏　名
	主 査	教授	鬼塚　真
	副 査	教授	原　隆浩
	副 査	教授	春本　要
	副 査	教授	藤原　融
	副 査	教授	松下　康之
	副 査	教授	下條　真司
	副 査	教授	George Fletcher

論文審査の結果の要旨

ソーシャルネットやウェブリンクなど多くの現象を表現できるグラフデータを処理することは広く関心を集めており，ノード分類はグラフ機械学習分野で最も重要な課題の一つである．グラフ深層学習（GNN）を含む機械学習手法はノード分類のための強力なツールである．しかし，それらの機械学習手法を実践的な応用に適用する場合，いくつかの課題が残されている．それらは，限定的なクラス構造の利用，限定的なスケーラビリティ，包括的な評価の欠落，限定的なデータタイプ，限定的な時系列のサポート，汎用事前学習モデルの欠落の六つである．

手法の有効性，スケーラビリティ，評価に関わる最初の三つの課題への解決策は，複数のノードやエッジタイプや時系列情報を含むグラフなどに代表される実践的な設定に拡張可能であり，他の全ての課題に対して基礎的な技術である．本論文では，これら三つの課題を解決する要素技術を提案している．主要な研究成果は次の通りである．

1) 手法の有効性を高めるためにクラス，属性およびグラフ構造の間の関係を捉える必要がある属性付きグラフにおけるクラスタリング問題（教師なしノード分類）に取組む．提案手法では，構造と属性の間の複雑な関係を捉えるために，構造と属性の間の非線形な射影関数を学習する．また，実データではエッジ情報が部分的にしか観測できないことをクラスタリングに考慮するために，positive unlabeled 学習を利用する．実験では提案手法が既存手法よりも高いクラスタリング精度を獲得することを示した．

2) GNN のスケーラブリティを改善するために，スケーラブルでない GNN を大規模グラフに適用可能な効率的でかつスケーラブルな GNN に変換するフレームワークを提案する．既存研究では個別のスケーラブルな GNN が提案されている一方で，提案フレームワークは多様な既存の GNN を効率的な事前計算ベースの GNN に変換する．また事前計算ベースの GNN をさらに効率化するために，グラフが持つ隣接行列と属性行列を GPU で計算可能な小さなブロックに分割するスキーマを提案する．実験では，入力されたスケーラブルでない GNN を提案フレームワークが精度を維持したスケーラブルな GNN に変換することを示し，提案スキーマが事前計算を単純な CPU での計算に比べ，約 2 倍高速化することを示した．

3) ノード分類手法の包括的な評価を可能とするために，多様な人工データを用いる評価フレームワークを提案する．まず，クラス，構造，属性間の関係を柔軟に捉えることができる属性付きグラフ生成器 GenCAT を開発する．次に GenCAT によって生成した多様なグラフデータを用いて，クラスサイズ，クラス間のエッジ割合，属性値，グラフサイズがそれぞれノード分類手法のパフォーマンスにどの程度影響するかに関して詳細な分析を行った．これらによって，既存研究では明らかになっていない近年の GNN が抱える課題を実証的に明らかにした．この知見は今後の GNN に関する研究に有益である．

以上のように，本論文は属性付きグラフにおけるノード分類に関する先駆的な研究として情報科学に寄与するところが大きい．よって本論文は博士（情報科学）の学位論文として価値のあるものと認める．