グラフニューラルネットワークを用いた
スプレッドシートの正規化

笹治　拓矢

政府や企業により，膨大な各種データが統計表データとしてウェブ上で広く公開されている．統計データとは，統計量である人口推移や工業製品の出荷額などの属性と値を組み合わせたデータである．これらの情報は政府や企業が調査・集計したものであり，一般に信頼度の高い情報源と考えることができる．そのため，多くの統計データを整理・分析することで知見を得て，仕事などに役立てられる．ただし，これらのデータは Excel 形式や CSV 形式による表形式データとして公開されており，中には見出しを階層関係にするなど複雑な階層関係を含むものも存在する．大量の統計データから知見を得るには，機械的な処理が安易となるデータに変換する必要があるため，複雑な見出し階層関係の認識タスクのほかに様々な正規化に関するタスクが提案されてきた．見出し階層の認識タスクにおける既存研究では，複雑な構造をもつ統計表を画像化し，表の局所的な特徴（罫線の有無や値間の位置関係）に基づいて階層関係を判定している．しかし，局所的な特徴による認識では，複数の罫線デザインで記述した見出しや，ツール上で罫線そのものを使用せずに図形を挿入する形で見出し階層を表現した場合に分類器が誤認識してしまう問題がある．そこで，スプレッドシート内の統計表を構成するセルの視覚的な情報（文字列や値，罫線など）をノード特徴とするグラフデータにグラフニューラルネットワーク (GNN) を適用し，スプレッドシートの大域的な特徴を考慮して統計表の見出しとその階層関係の認識を行う方法を提案する．また，教師データを用意するためには大きな労力が必要となるため，教師なし表現学習を行うことで認識タスクに有効なセル表現を学習し，少数の教師データであっても効果的な学習が可能となる方法を提案する．具体的には，教師なし表現学習手法の Deep Graph Infomax (DGI) を用いて有効な特徴表現を獲得する．我々の知る限りでは，本研究は見出し階層の認識タスクに GNN による教師なし学習手法を直接活用する最初の試みである．実験では，見出し階層のセルペアの認識と見出しセルの識別を対象とし，比較手法を従来の機械学習による手法として，事前学習済みの DGI から得られる埋め込みを用いて分類器の評価を行った．データセットには，政府統計ポータルサイト e-Stat 上から収集されたスプレットシートを使用し，認識タスクの正解ラベルを用意するためにアノテーションを行った．実験の結果，GNN の教師あり学習手法が見出し階層の認識タスクで最も精度が高くなり，見出しセルの識別タスクではノード初期特徴ベクトルを用いた分類器の精度が最も高くなることを示した．DGI の埋め込みは見出し階層の認識タスクのみ有効であるとわかった．このことから，GNN の教師あり学習手法は見出し階層の認識に効果的であり，見出しセルの識別ではノード初期特徴ベクトルを用いた従来の分類手法で十分であることが明らかになった．

（指導教員　　加藤　誠）