GCN による取引関係グラフからの企業の特徴量抽出

Feature extraction of companies from business relationship graphs by GCN

森 正和[1]　與五澤 守[1]　工藤 剛[2]

Masakazu Mori[1], Mamoru Yogosawa[1] and Tsuyoshi Kudo[2]

[1]株式会社日本総合研究所
[1]The Japan Research Institute, Limited
[2]株式会社三井住友銀行
[2] Sumitomo Mitsui Banking Corporation

Abstract: Financial institutions have a large amount of data on money transfer among companies. With these data, they can construct graphs that represent the business relationships of the companies. If they can predict a rating of a company's repayment capacity from the graphs, they can make more appropriate loan decisions and find new customers. In this study, we propose a method applying extended Graph Convolutional Networks to business relationship graphs representing the continuity of transactions to predict the ratings. As a result of applying this method to actual data, it was indicated that this method automatically extracted features necessary for the prediction of the rating.

１．はじめに

金融機関は企業に融資を行う際に，主にその企業の財務諸表（貸借対照表・損益計算書等）を参考にしてその企業の返済能力を算定し，融資実行の判断や融資額・利率等の決定を行う．しかし，中小企業や新興国の企業においては正確で信頼性の高い財務諸表を入手できない場合があり，結果的に適切な融資判断が行えない場合がある．また，金融機関がまだ融資したことがない企業の中から融資先として有望な企業を絞り込むために返済能力が低い企業を除外するには，事前に全ての企業の財務諸表が必要となるが，株式を公開していない企業の財務諸表は一般的に入手困難であり，従って返済能力が低い企業を除外することも困難である．

一方で，金融機関は企業の預金口座を保有しているため，企業間で取引がなされた際に発生する送金の情報（以下，この情報を「送金明細」と呼ぶ）を保有している．この送金明細から企業の返済能力を予測することができれば，金融機関が融資判断を行う際に企業から入手した財務諸表の信頼性を判定することが可能となる．また，金融機関が融資先として有望な企業を検索することも可能となる．

これまでデフォルト（債務不履行）の予測に関しては[１]や[２]をはじめ，多くの先行研究がなされてきたが，それらの先行研究はいずれも財務諸表を元にして作成した財務指標を説明変数としてデフォルトを予測するものである．本研究は財務諸表を使用しない点がこれらの先行研究と異なる．

また，財務諸表を使用しない先行研究として，[３]や[４]では入出金明細を元に数千の特徴量を設計・加工し，勾配ブースティング決定木等の機械学習モデルを用いてデフォルトの予測を行う手法が提案されている．本研究はこのような多くの特徴量を設計・加工する必要がない点が先行研究と異なる．

本研究では財務諸表や入出金明細の代わりに送金明細を使用する．送金明細は企業の取引関係を表しており，ここから企業をノード，企業間の取引関係をエッジとしたグラフを構築することが可能となる．近年，深層学習をグラフに応用した Graph Convolutional Network，及び，その派生手法が様々なグラフ分析タスクにおいて高い成績を収めている．（[５], [６], [７]）本研究ではこの Graph Convolutional Network を適用し，財務諸表に頼ることなく，企業の取引関係を表すグラフから自動的に特徴量を抽出し，企業の格付を予測する手法を提案する．

２．使用データと提案手法

２．１．使用データ

本研究で使用するデータは特定の一金融機関が保有する送金明細データと企業の格付データである．

送金明細データは振込・振替等の為替情報を元に

作成された企業間の送金情報データである．送金明細データは送金企業・受取企業・日付・金額の４つの情報が含まれる．企業が同じ金融機関に複数の預金口座を持っている場合や，他行にも預金口座を持っている場合には，企業名等の情報を元に同一企業として名寄せを行う．他行の預金口座間で行われた送金情報は送金明細データに含まれない．本研究では金額が 10 万円以上の送金明細のみを使用する．

企業の格付データは企業の返済能力を表す順序尺度を持ったカテゴリー変数である．格付データは各企業の決算期に作成された財務諸表等の情報を元に金融機関によって設定される．各企業の格付データは少なくとも 1 年に一度の頻度で見直されるが，債務不履行等特定のイベントがあった場合は随時見直される．格付が設定された企業は送金明細データに含まれる企業のごく一部である．

本研究では 3 年間の送金明細データを元に，その１年後に設定されていた格付が特定カテゴリー以下となっている返済能力が低い企業を予測する．なお，今回使用したデータにおいて，格付が特定カテゴリー以下の企業の割合は，訓練データは約 19%，テストデータは約 17%であった．

表１は本研究で使用した訓練データとテストデータの期間を示している．

表１：訓練データ・テストデータの期間

	送金明細	格付
訓練データ	2013 年 4 月～2016 年 3 月	2017 年 3 月時点の格付
テストデータ	2014 年 4 月～2017 年 3 月	2018 年 3 月時点の格付

２．２．取引関係グラフ

次に，送金明細データから取引関係グラフを作成する手順を説明する．

まず送金明細データを半期ごとに送金企業・受取企業を単位にして集計する．例えば，2013 年度上期に A 企業から B 企業に 100 万円が送金された，といった形で送金明細を集計する．

この半期ごとに集計されたデータを元に，格付の予測を行う企業（以下，「予測対象企業」と呼ぶ）ごとに，以下に示す手順で，販売先企業や仕入先企業（以下，両者をあわせて「取引先企業」と呼ぶ）との取引関係を示す有向グラフを構築する．

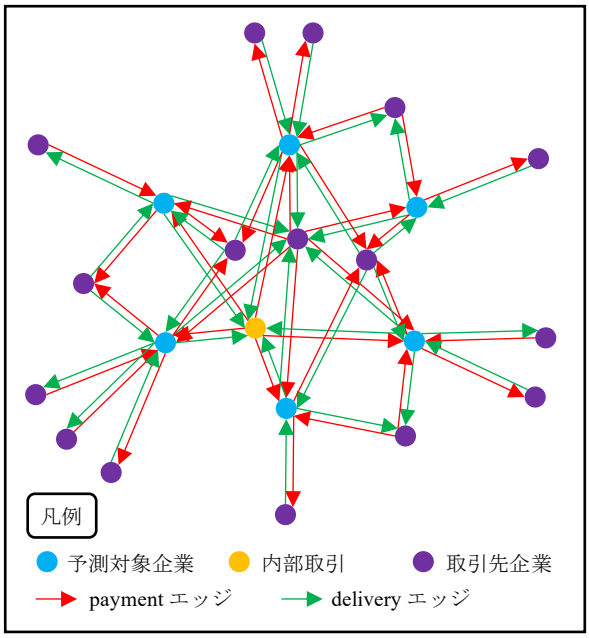
まず予測対象企業の３年間における各半期を示す６つのノードを作成する．次に，予測対象企業の内部取引を示すノードと取引先企業を示すノードを作成する．内部取引と取引先企業のノードは半期ごとに分割せず，３年間で共通の１つのノードとする．そして，予測対象企業の各半期に内部取引や取引先企業と取引が行われていた場合は，送金企業・受取企業間に双方向のエッジを結ぶ．送金方向のエッジを payment エッジ，逆方向のエッジを delivery エッジと呼ぶ．これらのエッジとは別に各ノードにはセルフループエッジを追加する．

ノード特徴量にはノードの種類（予測対象企業の各年度半期，及び，内部取引，取引先）を示す８次元のワンホットベクトル $\boldsymbol{h}_i^{(0)} \in \mathbb{R}^8\ (i = 1, \cdots, N)$ を設定する．ここで N は取引関係グラフのノード数を示す．

payment と delivery のエッジ特徴量 $p_{i,j} \in \mathbb{R}, d_{i,j} \in \mathbb{R}\ (i = 1, \cdots, N, j = 1, \cdots, N)$ には常用対数をとって 5 を引いた送金額を設定する．セルフループエッジにはエッジ特徴量を設定しない．

このようにして作成した取引関係グラフの例が図１である．この例にあるように，予測対象企業と継続した取引がある取引先は複数のエッジを持つのに対し，継続した取引が行われていない取引先はエッジが少なくなる．

図１：取引関係グラフの例



２．３．提案手法

取引関係グラフを入力として格付を予測するモデルを 4 層のニューラルネットワークで構築する．

まず 1 層目は Relational-GCN（[６]）を参考にして

下記に示す計算を行う． Relational-GCN との違いは正規化定数の代わりにエッジ特徴量を利用する点である．これによって，半期ごと・取引先ごとの送金額合計や受取額合計といった特徴量が内部的に生成される．

$$\boldsymbol{h}_i^{(1)} = \sigma\left(\sum_{j\in\mathcal{P}_{(i)}} W_p^{(0)} p_{j,i}\boldsymbol{h}_j^{(0)} + \sum_{k\in\mathcal{D}_{(i)}} W_d^{(0)} d_{k,i}\boldsymbol{h}_k^{(0)} + W_s^{(0)}\boldsymbol{h}_i^{(0)}\right)$$

ここで$\mathcal{P}_{(i)}$は i 番目のノードに向かって結ばれた payment エッジの始点ノード集合であり，$\mathcal{D}_{(i)}$は i 番目のノードに向かって結ばれた delivery エッジの始点ノード集合である．$W_p^{(0)}, W_d^{(0)}, W_s^{(0)}$は行数が隠れ層の次元数で列数が入力層の次元数（つまり，8）の重みパラメータ行列である．σは活性化関数であり，今回は ReLU 関数を用いた．

次に 2 層目は Graph Attention Network（[７]）を参考にして，下記に示す計算を行う．Graph Attention Network との違いはセルフループエッジとそれ以外のエッジで異なる重みパラメータを使用する点である．これによって，１層目で生成されたノードごとの特徴量を２層目に引き継ぎつつ，周辺ノードの特徴量を集約することが可能となる．

$$\boldsymbol{z}_i = W^{(1)}\boldsymbol{h}_i^{(1)}$$
$$\boldsymbol{s}_i = W_s^{(1)}\boldsymbol{h}_i^{(1)}$$
$$e_{i,j} = LeakyReLU\left(\boldsymbol{a}^{(1)^T}(\boldsymbol{z}_i||\boldsymbol{z}_j)\right)$$
$$f_i = LeakyReLU\left(\boldsymbol{b}^{(1)^T}\boldsymbol{s}_i\right)$$
$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k\in\mathcal{P}_{(i)}}\exp(e_{i,k}) + \sum_{l\in\mathcal{D}_{(i)}}\exp(e_{i,l}) + \exp(f_i)}$$
$$\beta_i = \frac{\exp(f_i)}{\sum_{k\in\mathcal{P}_{(i)}}\exp(e_{i,k}) + \sum_{l\in\mathcal{D}_{(i)}}\exp(e_{i,l}) + \exp(f_i)}$$
$$\boldsymbol{h}_i^{(2)} = \sigma\left(\sum_{j\in\mathcal{P}_{(i)}}\alpha_{i,j}\boldsymbol{z}_j + \sum_{k\in\mathcal{D}_{(i)}}\alpha_{i,k}\boldsymbol{z}_k + \beta_i\boldsymbol{s}_i\right)$$

ここで$W^{(1)}, W_s^{(1)}$は行数・列数が隠れ層の次元数となっている重みパラメータ行列である．$\boldsymbol{a}^{(1)}$は隠れ層の次元数の２倍，$\boldsymbol{b}^{(1)}$は隠れ層の次元数を持った重みパラメータベクトルである．||はベクトルの連結を示す記号である．この層ではエッジ特徴量$p_{i,j}, d_{i,j}$を用いない．

3 層目は取引関係グラフの予測対象企業各半期のノード特徴量，内部取引のノード特徴量，取引先のノード特徴量の平均値，最大値を連結したベクトルを入力とした全結合層とする．

$$\boldsymbol{h}^{(2)} = (\boldsymbol{h}_1^{(2)}||\boldsymbol{h}_2^{(2)}||\boldsymbol{h}_3^{(2)}||\boldsymbol{h}_4^{(2)}||\boldsymbol{h}_5^{(2)}||\boldsymbol{h}_6^{(2)}||\boldsymbol{h}_7^{(2)} || \underset{i\in\mathcal{N}_p}{\text{mean}}\boldsymbol{h}_i^{(2)} || \underset{i\in\mathcal{N}_p}{\max}\boldsymbol{h}_i^{(2)})$$
$$\boldsymbol{h}^{(3)} = \sigma\left(W^{(2)}\boldsymbol{h}^{(2)} + \boldsymbol{b}^{(2)}\right)$$

ここで$\boldsymbol{h}_1^{(2)}$から$\boldsymbol{h}_6^{(2)}$は予測対象企業各半期のノード特徴量である．$\boldsymbol{h}_7^{(2)}$は内部取引ノード特徴量である．$\mathcal{N}_p$は取引先ノード集合である．$W^{(2)}$は行数が隠れ層の次元数の９倍で列数が隠れ層の次元数となっている重みパラメータ行列である．$\boldsymbol{b}^{(2)}$は隠れ層の次元数を持ったバイアスパラメータベクトルである．

4 層目は予測対象企業の格付が特定のカテゴリー以下である確率を出力する全結合層とする．

$$h^{(4)} = \text{sigmoid}\left(\boldsymbol{w}^{(3)^T}\boldsymbol{h}^{(3)} + b^{(3)}\right)$$

ここで$\boldsymbol{w}^{(3)}$は隠れ層の次元数を持った重みパラメータベクトルである．$b^{(3)}$はバイアスパラメータである．

プログラムの実装においては PyTorch（[８]）と Deep Graph Library（[９]）を用いた．各隠れ層の次元数を 16 とし，損失関数にクロスエントロピー誤差関数を用いて確率的勾配降下法にて各重みパラメータ・バイアスパラメータの最適化を行った．この際，学習率は 0.001，オプティマイザーは Adam，バッチサイズは 100，エポック数は 20 とした．

実装したソースコードは下記 URL で公開している．

https://github.com/marisakamozz/feature_extraction_from_bizgraph

２．４．比較手法

提案手法との比較を行うため，勾配ブースティング決定木で同様の予測モデルを作成した．

勾配ブースティング決定木では取引関係グラフをそのまま入力データとして使用できないため，取引関係グラフを元に集計特徴量とグラフ構造特徴量を作成し，入力データとした．

取引関係グラフの集計特徴量は，半期ごと・予測対象企業ごと・送金または受取ごとに取引先企業の件数，合計金額，最小金額，最大金額，平均金額，

中央金額，金額の分散，金額の標準偏差を求め，特徴量とした．また，送金合計金額と受取合計金額の差額を追加の特徴量とした．

また，取引関係グラフのグラフ構造特徴量として，グラフ全体のノード数・エッジ数，及び，次数１から次数６までのそれぞれのノード数を特徴量とした．ここで次数の少ないノードの数は継続した取引が行われていない取引先企業数を，次数の多いノードの数は継続して取引が行われている取引先企業数を意図した特徴量である．

勾配ブースティング決定木の実装には LightGBM（[１０]）を用いた．訓練データを学習データと評価データに８対２で分割した上で，学習データの５分割交差検証と Optuna（[１１]）で最適なハイパーパラメータを決定した．

３．テスト結果と考察

３．１．テスト結果

表２は提案手法と比較手法を使ったテストデータに対する予測結果の AUC スコアと AR 値を示している．

表２：提案手法・比較手法のスコア

	提案手法（GCN）	比較手法（LightGBM）
AUC	**0.794**	0.776
AR 値	**0.588**	0.552

３．２．テスト結果の考察

提案手法の AUC スコアは比較手法に比べて高い値となった．これは取引関係グラフのグラフ構造に格付を判別するにあたって有益な情報が内包されている事，そして，Graph Convolutional Network がそのグラフ構造からうまく特徴量を抽出できていることを示唆している．

また，比較手法に対して提案手法が優れている点として，比較手法においては取引関係グラフに内包されている特徴量をデータサイエンティストが手作業で設計・加工しなければならないのに対し，提案手法ではニューラルネットワークが格付の判別に有益な特徴量を自動的に抽出可能である事があげられる．

３．３．規模に対する精度・再現率の考察

表３・表４は取引関係グラフのノード数・エッジ数に対する提案手法の精度と再現率を示している．取引関係グラフにおけるノード数・エッジ数は取引先企業の数や取引件数を表しているため，ここから企業の規模を推し量ることが可能である．

表３：ノード数に対する精度・再現率

ノード数	特定カテゴリー以下の割合	精度	再現率
~10	0.31	0.67	0.79
~100	0.20	0.80	0.48
~1000	0.07	0.93	0.07
1001~	0.01	0.99	0.00

表４：エッジ数に対する精度・再現率

エッジ数	特定カテゴリー以下の割合	精度	再現率
~10	0.32	0.64	0.81
~100	0.25	0.75	0.58
~1000	0.11	0.89	0.17
1001~	0.03	0.97	0.01

ノード数・エッジ数ともに 100 以下の企業については精度・再現率ともに高い値となっているが，ノード数・エッジ数が増えるに従って精度は高くなるものの再現率は低くなっていく．これは大規模な企業ほど特定カテゴリー以下の格付の企業が少ないため，モデルは大規模な企業を優良企業と予測しがちになることが原因と考えられる．そのため，提案手法は比較的規模の小さな企業に対して有効なモデルであると言える．

４．まとめ

本研究では，金融機関が保有する送金明細データから取引関係グラフを構築し，Graph Convolutional Network を用いて企業の格付を予測する手法を提案した．提案手法は財務諸表に頼らず企業の返済能力を予測するための手法である．

提案手法を実際のデータに適用した結果，取引関係グラフには格付を予測するために必要な情報が内包されており，Graph Convolutional Network によって自動的にその情報を抽出できることが示唆された．また，提案手法は比較的規模の小さな企業に対して有効であることが示された．

提案手法の課題としては，取引関係グラフは予測

対象企業の直接の取引先に関する情報しか利用できない点があげられる．取引先のさらにその先にある取引先に関する情報も利用することで，サプライチェーン全体の構造を取り込むことができれば，より正確に格付を予測することが可能となる可能性がある．

参考文献

[１] Altman, E. I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, The Journal of Finance, Vol. 23, No. 4, pp. 589–609 (1968)
[２] 高橋 久尚, 山下 智志:大規模データによるデフォルト確率の推定－中小企業信用リスク情報データベースを用いて－, 統計数理, Vol. 50, No. 2, pp. 241‐258 (2002)
[３] 本田 大悟，大古田 俊介，井實 康幸：入出金データを用いた企業デフォルト予測 ～機械学習手法の有効性比較評価～，第 48 回 2017 年度冬季 JAFEE 大会予稿集 (2018)
[４] 三浦 翔，井實 康幸，竹川 正浩：入出金情報を用いた信用リスク評価－機械学習による実証分析－，日本銀行ワーキングペーパーシリーズ，No.19-J-4 (2019)
[５] Thomas N. Kipf and Max Welling: Semi-supervised classification with graph convolutional networks, ICLR 2017 (2017)
[６] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov and Max Welling: Modeling Relational Data with Graph Convolutional Networks, arXiv:1703.06103 (2017)
[７] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò and Yoshua Bengio: Graph Attention Networks, ICLR208 (2018)
[８] PyTorch, https://pytorch.org/
[９] Deep Graph Library, https://www.dgl.ai/
[１０] LightGBM, https://github.com/microsoft/LightGBM
[１１] Optuna, https://optuna.org/