グラフエンベディングを活用した潜在取引関係予測

Latent Transaction Prediction Using Graph Embedding

藤塚 理史 [1]　工藤 剛 [2]

Masashi Fujitsuka[1], 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: Analysis of transaction data between corporations has strong future growth potential in finance since financial institutions have the large amount of the data. However, since it is difficult for each institution to get such data except the main customers' one, the data is partial and the application would be also limited. We set a similar environment artificially from the only observed data by removing some observed transaction links, and evaluate if we could predict the removed links. We show that graph embedding could lead to a solution of this problem.

１．はじめに

金融機関の主要な業務として顧客である企業の与信がある．この与信業務は，各企業の財務情報に基づいてしばしば行われるが，企業ごとのそれぞれの情報だけでなく，その企業がどのような企業と取引があるか，さらにその取引先企業はどのような企業と取引があるかといった取引関係データを活用することにより，従来の与信業務を高度化が出来る可能性がある．例えば，現状の与信判断における精度向上，システムによる自動化，また与信判断材料が不足している企業群に対する与信業務の適用拡大といったことなどが考えられる．取引関係と企業成長に関しても研究が行われており[1]，取引関係データの活用は大きく期待される．

しかし，取引関係データを活用する際の大きな課題として，全ての企業間の取引関係を把握することが出来ないことが挙げられる．例えば，1 つの金融機関では，その機関をメインバンクとしない企業の主要な取引関係を把握することは出来ないため，得られている取引関係のデータ量は多い一方で，多くの取引関係が未観測で欠損した状態になっている．そのため，取引関係データの活用範囲は限定的なものになってしまう．

上記のような課題に対して，観測出来ていないが実際には存在している取引関係（以下、潜在取引関係と記す）を予測することが出来れば，取引関係データの活用の幅を大きく広げることが可能となる．

そこで本研究では，潜在取引関係を擬似的に設定することにより，その潜在関係を予測する問題を考える．そして，そのような予測問題に対して，グラフエンベディング[2]と呼ばれる関係性データ（グラフデータ）から特徴を抽出する手法が有効であることを示す．

２．課題に対するアプローチ方法

問題設定

ここでのタスクは，潜在取引関係の予測を行うことであるが，潜在取引関係に当たるデータは入手が困難なため，実際にそのような関係を予測して評価を行うことは出来ない．そこで，次のように擬似的に潜在取引関係を設定することで，その関係を予測する問題に置き換える：

- ある企業に着目し，その企業が持つ取引関係をある割合（マスク率）でマスクする
- マスクされた取引関係を，潜在取引関係と見なし，それらの予測問題とする

これにより，未観測の取引関係を予測する問題を，現状保有するデータにおけるマスクされた取引関係を予測する問題と見なして考える．実際の課題と照らし合わせた場合，他の金融機関をメインバンクとするような企業ならば，マスク率が高い場合に相当すると考えられる．

グラフ分析としてのタスク

今回扱う取引関係データは，いわゆるグラフデータと呼ばれるデータ形式に対応する．グラフデータとは，ノードの集合 V と，リンクの集合 E の組 (V,E) からなる．各企業がノード，企業間取引がリンクに対応するグラフと見なすことが出来る．特に，今回の取引関係の予測は，グラフ分析におけるリンク予測の問題に対応する．

3．グラフ分析手法

3−1．リンク予測における指標

リンク予測とは，あるグラフデータが与えられた時に，それを手がかりにして，まだ知られていないリンクが存在するどうかを予測する問題である[3]．ノード u の隣接ノード集合を $\Gamma(u)$，隣接ノード数を $|\Gamma(u)|$ とする．ノードペア (u,v) に対して，リンクが存在するかどうかを表す指標として，下記のようなものが提案されている：

- 共通隣接ノード指標[4]

$$|\Gamma(u) \cap \Gamma(v)|$$

2 つのノードの間における共通の隣接ノード数を表したものである．つまり，共通の隣接ノードが多いノード同士はつながる可能性が高いことを表す．

- Jaccard 係数[5,6]

$$\frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

共通隣接ノード指標を，ノード u とノード v の隣接ノード全体で正規化したものである．つまり，2 つのノードの隣接ノードの全体の内，大半が重なるなら，つながる可能性が高いことを表す．

- Adamic/Adar 指標[7]

$$\sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}$$

共通隣接ノードの各ノードごとに異なる重みを設定し，足し合わせた指標である．例えば，どのノードともつながるようなノードとの共通リンクは，めったにリンクを持たないノードのリンクに比べて情報は多くないと考えられる．この指標は，そのような重みを考慮して設計された指標となっている．

- 優先的選択指標[8]

$$|\Gamma(u)| \cdot |\Gamma(v)|$$

この指標は，共通しているかどうかは関係なく，そもそも隣接ノード数が多ければ，リンクが存在しやすいとする指標である．例えば，どのノードともつながりやすいノードとは，そもそもつながる可能性が高いことを表す．

3−2．グラフエンベディング

上記で挙げたリンク予測の指標は，基本的に各ノードの隣接ノードの情報しか用いていない．一方で，今回の課題においては，隣接ノードの多くがマスクされて見えていないような状況であるため，このような指標は有効ではないように考えられる．そのため，隣接ノードの情報だけでなく，さらにそれより先の隣接関係性まで考慮することが必要である．そこで，そのような高次の隣接関係性まで考慮できるグラフエンベディングと呼ばれる手法をここでは考える．

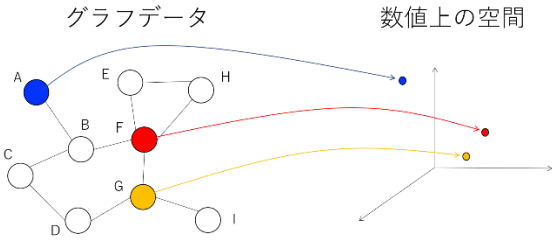

図 1：グラフエンベディングのイメージ

グラフエンベディングとは，グラフ上の性質を反映させるように，ノードごとに数値ベクトルを割り当てるものである．保ちたい性質に依存して，様々な手法が存在する．代表的な例としては，隣接ノードの関係性を保つために，以下のような最適化問題を解くことで，d 次元のエンベディング $f: V \to \mathbb{R}^d$ を求める Laplacian Eigenmaps[9]がある：

$$\min_f \sum_{(u,v)\in E} |f(u) - f(v)|^2.$$

近年，隣接ノード関係だけでなく，より高次の隣接関係性を保つようなグラフエンベディングの手法が複数提案されている[2,10-16]．ここでは特に，計算コストが少なく，柔軟にグラフの関係性を捉えることが可能な node2vec[2]と呼ばれるグラフエンベディング手法を用いる．node2vec は，ランダムウォークによるグラフ上の探索と，その探索された範囲における（広い意味での）隣接関係を保つようなエンベディングの学習からなる：

I.　ランダムウォークによるグラフ上の探索

グラフ上での性質を捉えるために，グラフ上でランダムウォークを行い，グラフデータをノードの系列データの集合に変換する．これにより，グラフ上での隣接関係性は，系列データにおける隣接関係性として置き換えられる．

この系列データにおける隣接関係性は，ランダムウォークに大きく依存する．node2vec では，ノード位置 u_{t-1} から u_t に移動した時，元の位置 u_{t-1} からの距離に基づいて，次に移動可能なノード x に対して，次のようなバイアスを設定する：

$$\alpha_{pq}(t,x)=\begin{cases}\dfrac{1}{p} & (if\ \ d(u_{t-1},x)=0)\\ 1 & (if\ \ d(u_{t-1},x)=1)\\ \dfrac{1}{q} & (if\ \ d(u_{t-1},x)=2)\end{cases}$$

ここで，$d(u,v)$ はノード u と v の距離を表す．パラメータ (p,q) がバイアスパラメータであり，例えば，p が q（かつ1）に比べて小さい場合，元いた位置に戻りやすくなる．逆に，大きい場合は，遠方に行きやすくなる．

このような探索により，単純な隣接ノードの関係性でなく，より広い意味でのグラフ上での隣接関係性を捉えられると期待できる．今回は，近傍，遠方に特に偏りがないように探索したいため，$p=q=1$ に設定する．

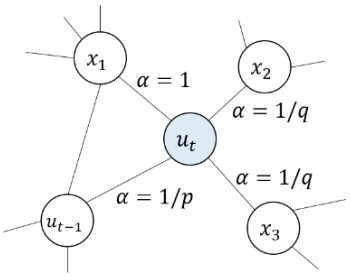

図 2：バイアスランダムウォーク

II.　探索範囲におけるエンベディングの学習

系列データからエンベディングを学習する手法として word2vec[17]を用いる．word2vec では，似たような文脈で出現する単語の表現は似ているとする分布仮説に基づき，単語列の前後で現れる単語を予測するモデル（Skip-gram モデル）を学習することで，エンベディングを得る[1]．この時の前後のコンテキストはウィンドウサイズを設定することにより柔軟に設定が可能である．

今回の場合は，ランダムウォークで得られた隣接関係性に基づいて，似たような関係性を持つノードが似たような表現になるようにグラフエンベディングが学習されることになる．これにより，3-1 に記載した隣接ノード関係だけを考慮した指標では捉えられないようなノード間の関係性が捉えられると期待できる．

４．実験

４−１．実験設定

ここでは，ある金融機関の 1 年間の企業間取引関係データを用いる．年間で送金、もしくは着金がある一定金額以上ある場合に、2 つの企業間に取引があるとして重みなし無向リンクを引く．このようにして作成した企業間取引関係グラフを今回の実験対象とする．企業数約 100 万社，約 1,000 万取引関係を含むグラフとなっている．

学習データ

企業を 1 社ランダムに選択し，その企業が持つ取引関係をあるマスク率でマスクする．ここでマスクされた取引関係が，正解となるテストデータに対応する．それに加えて，教師あり機械学習用の訓練データとして，ランダムに 10,000 個の取引関係を選択し，それらもマスクする．このようにして，マスクされたグラフデータが学習データに対応する．

評価データ

評価方法としては，上述でランダムに選ばれた 1 社に対して，正例（リンクが存在するノードペア）と負例（リンクが存在しないノードペア）の集合を与えた場合に，正例をいかに多く当てることが出来るかどうかの 2 値分類問題として評価を行う．正例は，上述のマスクした取引関係である．負例は，取引関係を持たない企業からランダムにサンプリングしたものに加えて，取引関係を持たない取引社数が多い企業からサンプリングしたものを同じ割合で用意して，それらを負例とする．

今回の対象企業は，図 3 のように取引社数が多い企業が指数関数的に減少するような分布となっている．そのため，ランダムにサンプリングすると，取引社数が少ない企業が多く抽出される．一方で，取引社数が多い企業とはリンクがつながりやすい傾向

[1] word2vec は，Skip-gram と Continuous Bag of Words（CBOW）の 2 つの定式化があるが，node2vec では前者を用いたモデルが採用されている．

があるため，正例の多くは取引社数が多い企業となりやすい．

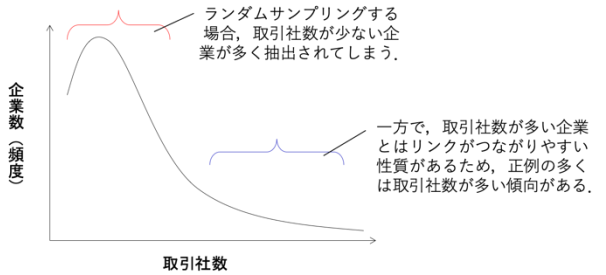

図 3：取引社数ごとの頻度分布（次数分布）

つまり，ランダムサンプリングを行うと，取引社数の大小によって，正例と負例が分かれる傾向となってしまう[2]．今回は，そのような傾向の違いを捉えられるかどうかを検証したいわけでないため，取引社数が多い企業から抽出した負例を加えたものを評価データとする．

評価方法

上述した評価は，1 つの企業に対する評価であり，サンプリングされた企業によって結果は大きく依存する．そのため，上述の評価をランダムにサンプリングした 20 社に対して独立に行い，その平均 ROC-AUC スコアで評価を行う．

具体的な評価としては，取引社数ごとに，100-300 社，300-600 社，600-1000 社の 3 つのカテゴリに分けて，それぞれで実施する．そして，それぞれの評価において，取引関係のマスク率を，20%，50%，80%とそれぞれ変えて評価を行う．

グラフエンベディング（node2vec）の学習

実装は，Stanford Network Analysis Project（SNAP）による C++実装を用いる[3]．上述したように，重みなし無向グラフとし，ランダムウォークのバイアスパラメータは，$p = q = 1$．探索は，各ノードを始点とした 80 長さのウォークを 10 セット実施．エンベディングのベクトル次元は 128，系列データの隣接ウィンドウサイズは 10 として学習を実施した．

4-2．比較手法

グラフエンベディング手法 node2vec を用いた場合と，3-1 に記載した 4 つの指標（以下，ベースライン指標と記す）との比較を行う．node2vec に関しては，2 つのノードから得られるそれぞれのエンベディングベクトル $f(u), f(v)$ から，cos 類似度，

$$\frac{f(u) \cdot f(v)}{|f(u)|\ |f(v)|}$$

を算出することで指標として用いる．

また，教師あり機械学習の評価として，

- ベースライン指標を特徴量としたモデル
- ノードペアそれぞれの node2vec で得られるエンベディングの要素積を取り，各次元をそれぞれ特徴としたモデル[4]

の 2 つを評価する．教師あり機械学習モデルとしては，勾配ブースティングモデルを使用する．学習用の正例データは，4-1 の学習データ作成の際に除去した 10,000 取引関係，負例データは，取引関係がない企業ペアからランダムにサンプリングした 10,000 取引関係を用いる．

4-3．実験結果と考察

表 1 に実験結果を記載した．ベースライン指標に比べて，高い精度で予測出来ていることが分かる[5]．特筆すべきは，マスク率を上げた場合の，精度の減少度合いである．ベースライン指標では，およそ 0.05 程度減少している一方で，グラフエンベディングを用いた場合，その半分程度である．これは，保有している取引データが部分的なものである場合でも，グラフエンベディングを用いることで取引関係を予測出来ることを示唆している[6]．

ベースライン指標は，隣接ノードの関係性しか見ていないため，マスク率が高くなるに従って，その予測精度は急激に減少してしまっていると考えられる．一方で，グラフエンベディング（node2vec）は，マスク率が高い状況でも，その先の隣接関係性まで捉えていると考えられるため，減少度合いが少なく，有効に機能すると考えられる．

---

[2] この場合，リンク数が多いほど高い確率とする優先的選択指標が最も有効となる．

[3] https://github.com/snap-stanford/snap/tree/master/examples/node2vec

[4] node2vec を用いたリンク予測の特徴量としては，要素積を取る以外の方法もあるが，[2]におけるリンク予測の実験で要素積による特徴が最も精度が高い結果であったことから，今回は要素積を選択する．

[5] 優先的選択指標に関しては，相手先の企業の取引数にしか依存しないため，マスク率よる影響はほとんど受けない結果となる．

[6] 今回の結果は，ランダムに選択された 20 社の平均精度である．1 つ 1 つの結果で見た場合，ベースライン指標の方が高いケースもいくつか存在する．

取引社数範囲	マスク率	教師なし学習					教師あり学習	
		ベースライン指標				グラフエンベディング cos 類似度	ベースライン指標	グラフエンベディング
		共通隣接ノード指標	Jaccard 係数	Adamic/Adar 指標	優先的選択指標			
100-300 社	20%	0.662	0.679	0.671	0.565	**0.796**	0.684	**0.843**
	50%	0.657	0.663	0.665	0.563	**0.794**	0.666	**0.844**
	80%	0.613	0.601	0.620	0.565	**0.773**	0.617	**0.815**
300-600 社	20%	0.657	0.695	0.673	0.523	**0.782**	0.690	**0.838**
	50%	0.639	0.673	0.651	0.526	**0.770**	0.656	**0.824**
	80%	0.606	0.622	0.614	0.525	**0.760**	0.594	**0.816**
600-1000 社	20%	0.644	0.671	0.652	0.518	**0.765**	0.679	**0.802**
	50%	0.635	0.670	0.642	0.514	**0.750**	0.655	**0.798**
	80%	0.588	0.616	0.592	0.512	**0.742**	0.577	**0.783**

表 1：実験結果（20 社の平均 ROC-AUC スコア）

残された課題

今回の評価方法は，着目した企業に対して，取引関係をあるマスク率でマスクするというような特定の状況を考えたが，現実的には，ある金融機関をメインバンクとしない企業は複数存在するため，マスクの取り方をより複雑にした設定における評価が必要になると考えられる. また, リンク予測の性質上, 取引社数が少なすぎると, うまく機能しない. また, 多すぎる場合も問題が困難になると考えられるため, 適用可能範囲を明確にすることも求められる.

また，今回は企業ごとのつながりの情報しか使っていないが, 企業が持つ属性情報も活用することで, より高度な予測につなげられると考えられる．例えば，今回の分析では，口座を持たない企業，つまりグラフ内にノードが存在していないものは自動的に対象外とするが，属性情報用いることで企業間の類似度の算出が可能となり，全く関係性を持たないような企業に対してもアプローチが可能となるといった方向性も期待出来る.

５．結論

本研究では，取引関係データが部分的にしか得られないという課題に対して，観測出来ていないが実際には存在する取引関係（潜在取引関係）を保有している取引関係データだけから疑似的に設定することにより，それらを予測する問題として定式化し，その問題に対し，グラフエンベディングと呼ばれる手法が有効であることを示した．これにより，現状保有しているデータ量だけでは見えない部分の補完が可能となることから，取引関係データの活用の幅を大きく広げることが出来ると考えられる.

参考文献

[1] みずほ情報総研株式会社，平成２８年度 中小企業の成長に向けて中長期的に 取り組むべき施策の検討に向けた 我が国中小企業の成長過程分析に係る委託調査, 経済産業省平成28年度委託調査報告書, 2018.

[2] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[3] 鹿島久嗣，ネットワーク構造予測，人工知能学会誌 Vol.22 No.3, 2007.

[4] Newman, M. E. J.: Clustering and preferential attachment in growing networks, Physical Review Letters E, Vol. 64(025102), (2001)

[5] Baeza-Yates, R. A. and RibeiroNeto, B. A.: Modern Information Retrieval, A 8 Press / Addison-Wesley (1999)

[6] Liben-Nowell, D. and Kleinberg, J.: The Link Prediction Problem for Social Networks, in Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM), pp. 556–559 (2004)

[7] Adamic, L. A. and Adar, E.: Friends and neighbors on the Web, Social Networks, Vol. 25, No. 2, pp. 211– 230 (2003)

[8] Barab´asi, A. L., Jeong, J., N´eda., Z., Ravasz, E., Shubert, A., and Vicsek, T.: Evolution of the social network of scientific collaborations, Physica A, Vol. 311, No. 3-4, pp. 590–614 (2002)

[9] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In NIPS, 2002.

[１０] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701–710. ACM, 2014.

[１１] Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A.J. Smola. Distributed large-scale natural graph factorization. In WWW, 2013.

[１２] S. Cao, W. Lu, and Q. Xu. Grarep: Learning graph representations with global structural information. In KDD, 2015.

[１３] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu. Asymmetric transitivity preserving graph embedding. In KDD, 2016.

[１４] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In WWW, 2015.

[１５] S. Cao, W. Lu, and Q. Xu. Deep neural networks for learning graph representations. In AAAI, 2016

[１６] D. Wang, P. Cui, and W. Zhu. Structural deep network embedding. In KDD, 2016.

[１７] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.