

頂点間の類似度の足し合わせによるリンク予測精度の改善

元田 剛史[†] 村田 剛志^{††}

東京工業大学 大学院情報理工学研究科 計算工学専攻

1. はじめに

近年、科学の様々な分野においてネットワークで表現されるデータの解析が盛んに行われている。データの解析の一つに、ネットワークの観測部分から未観測部分の構造を予測するリンク予測問題がある。リンク予測問題は、タンパク質間の未観測な相互作用の予測や、ソーシャルネットワークにおける新たな人間関係の予測など、様々な研究やサービスの発展に寄与するものと期待されている。これまで様々なリンク予測手法が提案され、その中でもノードペアの類似度を用いたリンク予測手法は盛んに研究されている手法である。類似度を用いたリンク予測手法は計算が高速であるが、一部のネットワークに対して精度が低くなることが知られている。その為類似度を用いず、最尤推定を用いてネットワークの生成モデル (HRG) を学習しリンクを予測する手法 [Clauset 08] も提案されている。Clauset らの手法は高い予測精度を達成可能であるが、計算時間が問題となっている。本論文では多種類の類似度 (Adamic/Adar (AA), Shortest Path (SP), Preferential Attachment (PA)) を足し合わせることで、高速計算可能でかつ、高精度達成可能なリンク予測手法を提案する。また、多数の分野から網羅的に収集した 28 の実ネットワークに対して、提案手法と AA, SP, PA, HRG の予測精度を比較した。その結果、提案手法は HRG より高速かつ、17 のネットワークに対して、HRG より高い予測精度を達成した。ここではリンク予測問題を、ノードペアをリンクのありそうなものから降順に並べるランキング問題と定義し、ランキングより算出される AUC (Area Under ROC Curve) をリンク予測精度とする。

2. 従来のリンク予測手法

[Getoor 05] によると、リンク予測問題へのアプローチは 2 種類に大別できる。一方は、[Liben-Nowell 04] らのノードペアの類似度を用いたリンク予測手法のようにネットワークの構造的性質を用いる方法であり、他方は属性情報を用いる方法である。また、それら両者を組み合わせた手法も提案されている。しかし、いずれを用いても高い精度でのリンク予測は難しく、またネットワークのデータに属性情報が付与されていない場合も多い。そこでリンク予測の精度を改善するため、各リンクの有無をノードペアごとと独立に予測するのではなく、観測されたネットワークからネットワーク全体のリンク構造を予測する方法が考案された。[Clauset 08] らは任意の階層構造を持つネットワークを生成できる Hierarchical Random

Graph (HRG) と呼ばれる汎用モデルを定義し、HRG を用いたリンク予測手法を提案している。HRG はパラメータを調整することにより観測部分のネットワークの生成モデルを定義し、その生成モデルを用いて未観測部分のリンクを予測する。

類似度を用いたリンク予測手法

類似度を用いたリンク予測手法は、ネットワークにおいて類似度の高いノードペアに新しいリンクが生まれやすいと仮定し、類似度に基づいてノードペアをランキングすることでリンク予測問題を解決する。[Lu 10] らはこれまで提案されてきた多数の類似度を分類している。提案手法では Lu らの分類、及び各種リンク予測手法の適用範囲に関する研究の成果 [元田 09] から、性質が異なり、かつ高速計算可能な類似度 3 種類を組み合わせに用いた。以下にそれら 3 種類の類似度を紹介する。

Adamic/Adar (AA)

共通隣接ノード数が多いノードペアを上位にランキング。共通隣接ノードに着目している為、距離 2 で離れているノードペアに対してのみ類似度を定義。

Shortest Path (SP)

ネットワーク上での距離が近いノードペアを上位にランキング。連結成分のノードペアに類似度を定義。

Preferential Attachment (PA)

次数の大きいノードペアを上位にランキング。ネットワーク上での距離に関係なく次数 1 以上の全てのノードペアに類似度を定義。

3. 提案手法について

提案手法では AA, PA, SP の類似度を重みつき線形和を用いて足し合わせ、新たな類似度を計算する。まず足し合わせにあたり、リンク予測を行うネットワークごとに高精度を達成可能な重み教師あり学習 (ロジスティック回帰) を用いて学習する。重みを学習後、入力ネットワークから各ノードペア間の類似度 (AA, SP, PA) を計算し、学習した重みを用いて足し合わせることで新たな類似度を計算する。以下に提案手法における重みの学習法を紹介する。

重みの学習

入力ネットワークに対して、リンク予測の精度が高くなると考えられる重みを図 1 に従って学習する。図 1 において、①では②~③の教師あり学習で用いるデータを作成している。入力データからランダムにリンクを取り除き、取り除いたリンクを教師データ、残ったリンクを訓練データとする。ただし、ランダムの影響を軽減する為に訓練データと教師データの組を k 個生成し、③において bagging、又は 10-fold cross validation (10-fold c.v.) を用いてモデル選択を行う。教師あり学習では、まず②で訓練データから全ノードペアの類似度 (AA, SP, PA) を計算し、次に③で教師データを高精度で予

Improving the accuracy of link prediction by combining similarities of node pairs

[†]Takeshi Motoda (motoda@ai.cs.titech.ac.jp)

^{††}Tsuyoshi Murata (murata@cs.titech.ac.jp)

Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

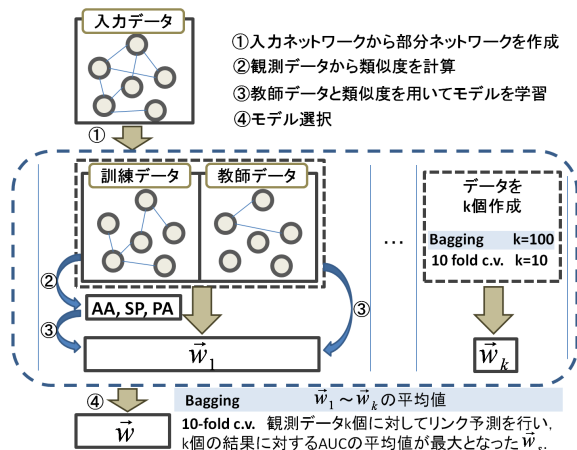


図 1: 提案手法における，入力ネットワークごとに適切な重みを学習する方法についての概要図．

測できるような，類似度を足し合わせる際の重みを学習する．このような重みを学習する問題は，各ノードペアの類似度を特徴量としたロジスティック回帰問題として定義でき，最尤推定法を用いてロジスティック回帰問題を解くことで，重みを学習している．

また②～③では，訓練データと教師データのペアごと独立に計算できる為，並列計算を用いることにより更なる計算の高速化が可能である．

4. 実験

本論文では提案手法の有用性を確認するため，Web 上から集めた 28 の実ネットワークに対して実験を行い，6 種のリンク予測手法 (AA, SP, PA, 提案手法 (bagging), 提案手法 (10-fold c.v.), HRG) のリンク予測精度を計測した．

4.1 実験データ及び実験環境

実験には HRG の適用範囲 [Clauset 08] に合わせてノード数 1000 以下のネットワークを用い，全て無向グラフとして扱った．実験には CPU: Intel (R) Core (TM2 Quad CPU Q9400 (2.66GHz))，RAM: 4GB 搭載 PC を用い，提案手法は重みを学習する際 4 コアを用いて並列計算を行っている．

4.2 実験方法

本実験では入力ネットワークからリンクをランダムに 1 割～5 割取り除き，残ったネットワークから各リンク予測手法を用いて取り除いたリンクを予測する．この際，ランダムの影響を軽減する為，取り除くリンクの割合ごとに部分ネットワークを 100 個用意する．各手法から得られたランキングに対して，取り除いたリンクを正例，残りのリンクを負例とした AUC を計算し，各割合 100 個の実験結果に対して，手法ごとに AUC の平均をとる．ここで得られた，手法ごとの AUC の平均値を各手法のリンク予測精度とする．

4.3 実験結果

4.3.1 リンク予測精度について

提案手法 (bagging) 及び提案手法 (10-fold c.v.) は非常に近い予測精度を達成するが，提案手法 (bagging) は 28 の実ネットワーク中 16 のネットワークに対して 提案手法 (10-fold c.v.) より高い予測精度を達成した．以下では 2 つのモデル選択の方法をまとめ，提案手法として結論を述べる．提案手法はリンク数 100 以上の多くのネットワークに対して，HRG より高いリンク予測精度を達成した．とりわけ，AA, SP, PA

ネットワークの種類	総ネットワーク数	高精度を達成したネットワーク数
人間関係	13	9
引用・共著	2	2
辞書	2	2
生体	3	3
交通網 (陸路)	1	1
交通網 (空路)	1	1
貿易	3	0
論理回路	2	0
食物連鎖	1	0

表 1: 実験結果「総ネットワーク数」は実験に用いた各種類の実ネットワーク数を表し，「高精度を達成したネットワーク数」は提案手法 (bagging) が HRG より高精度を達成したネットワーク数を表す．

のうち何れかが高精度を達成するネットワーク [元田 09] に対しては高い予測精度を達成した．また (AA, SP, PA) が苦手とするネットワーク [元田 09] (貿易ネットワーク，論理回路ネットワーク，食物連鎖ネットワーク) に対しては，(AA, SP, PA) が達成した精度のうち，最高の精度と同等の精度を達成した．これらのネットワークに対しては，組み合わせに用いる類似度を増やす事で予測精度を改良できるものと考えられる．

4.3.2 計算時間について

提案手法はモデル選択に用いた方法に関わらず，共に HRG より高速に計算を終了した．American college football* に対して，提案手法 (bagging) は HRG の約 1/2 の実行時間で計算を終了し，提案手法 (10-fold c.v.) は HRG の約 1/10 の実行時間で計算を終了した．

5. おわりに

本論文では性質の異なる複数の類似度を組み合わせる事により，HRG より高速な手法を提案し，28 の実ネットワークに対する実験の結果，17 の実ネットワークに対してその有用性を確認した．本手法はネットワークごとに重みを学習し，最適な類似度を用いてリンクを予測するため，様々な種類のネットワークに対し高精度なリンク予測が可能である．また，本手法は組み合わせに用いる類似度の数を増やすことにより，更なる精度改善が可能と考えられる．加えて，本手法は並列計算を用いることにより計算の高速化が可能である．

参考文献

- [Clauset 08] Clauset, A., Moore, C., and Newman, M. E. J.: Hierarchical structure and the prediction of missing links in networks, *Nature*, Vol. 453, pp. 98–101 (2008)
- [Getoor 05] Getoor, L. and Diehl, C. P.: Link Mining: A Survey, *SIGKDD Explorations*, Vol. 7, No. 2, pp. 3–12 (2005)
- [Liben-Nowell 04] Liben-Nowell, D. and Kleinberg, J.: The Link Prediction Problem for Social Networks, In *Proc. of the 12th Int. Conf. on Information and Knowledge Management (CIKM)*, Vol. 67, pp. 556–559 (2004)
- [Lu 10] Lu, L. and Zhou, T.: Link Prediction in Complex Networks: A Survey, *Physics and Society (physics.soc-ph); Social and Information Networks (cs.SI); Computational Physics* (2010)
- [元田 09] 元田 剛史, 村田 剛志, 佐藤 泰介: 実ネットワークに対する，各種リンク予測手法の適性について，2009 年度人工知能学会 全国大会 (第 23 回) 論文集, Vol. 3B1, No. 1, pp. 4–8 (2009)

* Network data

(<http://www-personal.umich.edu/~mejn/netdata/>)