論 文

Technical Papers

ネットワーク構造の確率的な時変モデルに基づく教師ありリンク予測

A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction

鹿島 久嗣
Hisashi Kashima
日本アイ・ビー・エム株式会社 東京基礎研究所
Tokyo Research Laboratory, IBM Research
hkashima@jp.ibm.com

安倍 直樹
Naoki Abe
アイ・ビー・エム株式会社 T.J. ワトソン研究所
T.J.Watson Research Center, IBM Research
nabe@us.ibm.com

keywords: link prediction, link mining, network evolution model, biological network

Summary

We introduce a new approach to the problem of link prediction for network structured domains, such as the Web, social networks, and biological networks. Our approach is based on the topological features of network structures, not on the node features. We present a novel parameterized probabilistic model of network evolution and derive an efficient incremental learning algorithm for such models, which is then used to predict links among the nodes. We show some promising experimental results using biological network data sets.

1. はじめに

近年, ネットワーク構造をもったデータの解析の必要性が高まっている. SNS(ソーシャルネットワーキングサービス) の流行によってにわかに注目を集めている社会ネットワークは個人や企業などの社会的主体と, 友人関係などのそれらの間のの関係によって表現されており, その構造は一般的にネットワーク構造 (グラフ) 構造によって表現することができる. バイオインフォマティクスの分野では, 遺伝子やタンパク質といった要素の間のネットワーク構造は, 制御関係や物理的な相互作用を表している (図 1). 要素間のリンクは静的な関係にとどまらず, 時間とともに移り変わることもありうる. たとえば, 電子メールの交換や協業などは一時的な関係であるといえる. このように多くの要素が複雑に関わりあうシステムをモデル化し, 解析するためには, 個々の要素のみに注目するだけでは十分とはいえない. システム全体としての性質は, しばしば相互作用や因果関係などといった要素間の関係のなかに埋め込まれている. 社会計量学の分野においては, このような観点から社会ネットワークの解析に関する研究が長年行われている [Wasserman 94].

一方, 近年データマイニングの分野では**リンクマイニング** [Getoor 05] などと呼ばれ, 従来の社会解析で対象とされてきたデータよりも遥かに大きいネットワーク構造データに対してさまざまな解析が試みられるようになってきた. リンクマイニングのタスクは, リンク構造を用いた要素の分類／ランキング／クラスタリング, リンク予測, 部分グラフ発見などをはじめ多岐にわたる [Getoor 05]. これらの中でも, 本論文では, 観測されたネットワークの部分から, 残りの部分を予測する (あるいは, 現在のネットワーク構造から, 将来のネットワーク構造を予測する) 問題であるリンク予測問題をとりあげる. リンク予測問題の応用としては, 例えば社会ネットワークにおける人間間の友人関係や上下関係, あるいは, 交流や協業などの将来の振る舞いを予測したり, バイオインフォマティクスにおいては, タンパク質間の相互作用や, 制御関係を予測し, 新たな生物学的知識を発見するための実験の手がかりを見つけたりなどといった使い方が期待できる.

リンク予測の問題は通常, ノードのペアについての 2 値分類問題あるいはランキング問題として定式化することができる. この際, 予測に用いることのできる情報としては 2 種類あり, ひとつはノード属性 (ノード自身の持つ情報) であり, もうひとつは構造属性 (ネットワーク構造の持つ情報) である. ノード属性に基づくリンク予測の試みはいくつか存在するが, 構造属性に基づくアプローチは現在までのところあまり行われておらず [Popescul

03, Hasan 05, Taskar 03, O'Madadhain 05], 本論文では後者に着目することにする．構造属性は通常, ネットワーク構造の生成モデル (例えば Barabási ら [Albert 99] の preferential attachment モデルなど) から導かれ [Newman 03], Liben-Nowelly と Kleinberg は各種のネットワーク生成モデルなどから導かれる指標を比較している [Liben-Nowelly 04].

本論文で我々は, パラメトリックなネットワーク構造の変遷モデルを提案し, これをリンク予測に用いることを提案する．提案手法は, 学習によって調整可能なパラメータを持つことで, リンク予測性能の向上を期待できる点で従来のものとは異なっている．我々の提案するモデルは, リンクの存在／非存在が確率的に反転するような「コピー&ペースト」の機構に基づいており, このモデルによる平均的なネットワーク構造が定常状態にあることを仮定することで, このモデルに基づくリンク予測を行うことができる．定常状態の推定はトランスダクションの問題として捉えることができるが, 我々は, 未知の構造の推定と, 指数勾配乗算 (exponentiated gradient) 型 [Helmbold 95] に基づくパラメータ推定を交互に行うトランスダクション手法を提案する．

最後に, 代謝ネットワークやタンパク質相互作用ネットワークなどの生体ネットワークデータを用いた実験において, 提案手法によって, 幾つかのリンク予測指標を上回る性能が得られることを示す．

本論文の構成は以下のとおりである．まず 2 章においてリンク予測問題を定義する．3 章ではネットワーク構造変遷の確率モデルを提案し, 4 章でこれに基づくリンク予測の方法を提案する．5 章では関連研究を述べ, 6 章では計算機実験の結果を述べる．7 章は結論および今後の研究とする．

2. リンク予測問題

まず, この論文で扱うリンク予測問題を定義する．対象とする構造ネットワークデータを $G=(V,\phi)$ とする．ここで, $V=\{1,2,\ldots,|V|\}$ はノードのインデックス集合, $\phi: V\times V\to[0,1]$ は後述する**リンクラベル**関数とする．各ノード $v\in V$ はネットワークの構成要素, たとえば SNS における 1 人の参加者, タンパク質相互作用ネットワークにおける 1 つのタンパク質を表している．**リンクラベル** $\phi(i,j)$ は, ノードペア (i,j) の間にリンクが存在するか否かを表し, リンクが存在するなら $\phi(i,j)=1$, しないなら $\phi(i,j)=0$ となる．ϕ は対称であることに注意する．

$E^L\subset V\times V$ を枝ラベルが分かっているノードペアの集合とし, これを**ラベル付けされたペア**と呼ぶ．**リンク予測問題**は, V と E^L が与えられたときに, 残りのノードペア $E^U:=(V\times V)-E^L$ に対してリンクラベルを予測する問題である．例えば, SNS において友人を推薦する機能や, タンパク質相互作用ネットワークにおいて未知の相互作用を発見するような問題がこれにあたる．

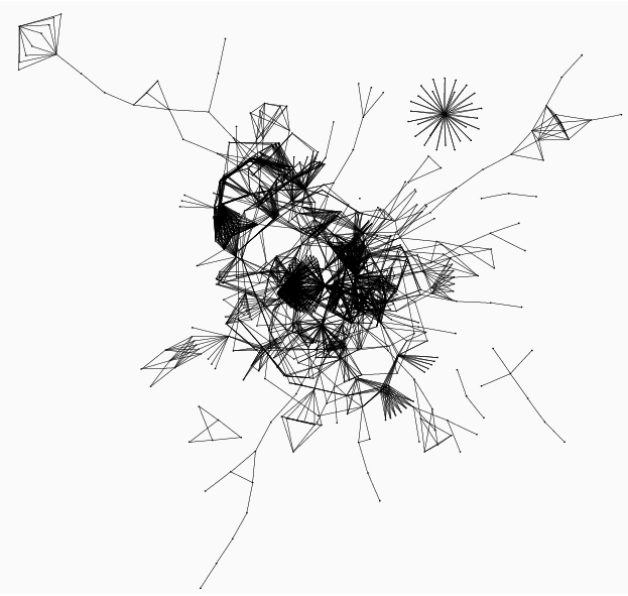


図 1 S. Cerevisiae の代謝ネットワーク．ノードはタンパク質を, リンクは連続する反応が 2 つのタンパク質によって行われることを示している．

本論文では, いくつかの既存研究で用いられているような, 各ノードのもつ情報 (人間の場合, 名前や年, 趣味などといった情報) であるノード属性は用いない．勿論, 最終的には, 全ての入手可能な情報を全て用いることが望ましいが, 本論文ではネットワーク構造の持つ情報である構造属性に注目し, これのみに基づくリンク予測問題を考えることにする．

3. ネットワーク構造の変遷モデル

この章では, ネットワーク構造の時間的変遷を捉えたパラメトリックな確率モデルを提案する．尚, 「ネットワークの構造」といったとき, 任意のノードペアについてのリンクの有無を示すリンクラベル関数 ϕ と同義であるとする．$\phi^{(t)}$ を時間 t におけるリンクラベル関数とする．ここで, $\phi^{(t)}$ は時間とともに変化するが, ネットワークの構成要素 V は変化しないものとする．

ϕ の値が確率的に反転するようなモデルを考えよう．反転は全く一様にランダムに起こるわけではなく, ネットワークのもつ何がしかの特徴に基づいて起こるものとする．我々はこの機構を, ノード間での「枝のコピー&ペースト」によって特徴付ける．コピーの起こる確率は, ネットワークの特徴を反映してノードごとに異なるものとする．また, 我々はネットワーク構造の変化にマルコフ性を仮定する, つまり, $\phi^{(t+1)}$ は $\phi^{(t)}$ のみに基づいて決まるものとする．

提案するモデルでは, ある時点において, ノード ℓ から

ノード m に確率 $w_{\ell m}$ でコピーが起こるものとする．したがって, 次のような確率的な制約が存在する．

$$\sum_{\ell m} w_{\ell m} = 1,\ w_{\ell m} \geq 0. \tag{1}$$

尚, $w_{\ell\ell} = 0$, すなわち各ノードは自分自身へのリンクラベルのコピーは行わないとする．また, W を (ℓ, m) 番目の要素が $w_{\ell,m}$ であるような行列とする．一旦ノード ℓ からノード m にコピーすることが決まると, 今度はノード ℓ のもつ $(\phi^{(t)}(\ell, m)$ を除いた$)|V|-1$ 本のリンクラベルの中から一様にランダムに枝を選び, それをノード m に対してコピーする．

このモデルの背景にあるのは次のような考え方である．ある人 A が, 他の人 B に対し大きな影響力を持っている場合, A の交友関係は B の交友関係に大きく影響するであろう．また, 生物の進化の過程においては遺伝子のコピーが起こることがあるが, この場合も新しくコピーされてできた遺伝子は, もとの遺伝子の性質の大部分を受け継ぐであろう．この仮説に従えば, ノード k がノード i に大きな影響力をもち, かつ, ノード k とノード j の間にリンクが存在するとき, ノード i とノード j の間には恐らくリンクが存在するだろうと結論づけることができ, 反対に, ノード k とノード j の間にリンクが存在しない場合には, ノード i とノード j の間には恐らくリンクが存在しないだろうと結論づけることができる．

ある時点 t において, $\phi^{(t)}(i,j)$ が特定のリンクラベルをもつ場合, その理由は 2 つ考えられる．1 つはあるノード k から i または j へのコピーが起こった場合であり, もう 1 つは, どこか別の場所でコピーが起こったため, もともとのリンクラベルが変化せずそのままであった場合, すなわち $\phi^{(t)}(i,j) = \phi^{(t+1)}(i,j)$ である場合である．

この議論に基づき, 時点 $t+1$ においてノード i とノード j の間にリンクが存在する確率は次のように書くことができる．

$$\begin{aligned}\Pr[\phi^{(t+1)}(i,j) = 1] = \\ \frac{1}{|V|-1}\Big(\sum_{k\neq i,j} w_{kj}\phi^{(t)}(k,i) + w_{ki}\phi^{(t)}(k,j)\Big) \\ +\Big(1 - \frac{1}{|V|-1}\sum_{k\neq i,j} w_{kj} + w_{ki}\Big)\phi^{(t)}(i,j).\end{aligned} \tag{2}$$

ここで第 1 項目は, ノードペア (i,j) に対するリンクラベルがコピーによって書き換えられる確率を指し, 第 2 項目は, 書き換えられずに (コピーが別の箇所で起こったため) そのままであった確率を指している．第 1 項目で, ノード k が, ノード i に対してノード j との間にリンクを持つようにさせるためには, ノード k は時点 t においてノード j との間にリンクを持っていなければならない $(\phi^{(t)}(k,j) = 1)$. 同様に, ノード j に対してノード i との間にリンクを持つようにさせるためには, ノード k は時点 t においてノード i との間にリンクを持っていなければならない (i.e. $\phi^{(t)}(k,i) = 1$). また, ノードペア (i,j) に対するリンクラベルは i から j, あるいは j から i にはコピーされないものとする．

なお, (2) は, 周辺確率を評価しており, つまり, 各リンクラベルの確率を独立に評価している．従って, (2) の積は全てのリンクラベルの同時確率にはなっていないことに注意する．

4. リンク予測

この章では, 前の章で定義したネットワーク構造の変遷モデルを用いて, 部分的に観測されるネットワーク構造からモデルのパラメータを推定するとともに, 残りの構造を予測する手法を提案する．

4·1 ネットワーク構造変遷モデルの定常期待状態

モデル (2) において, 真のパラメータは未知であり, また, 問題によってはネットワーク構造変化を時間的に追ったデータは得られない (例えば, タンパク質相互作用ネットワークなど) ためリンク予測を行うのは不可能である．そこで我々は, 「現在のネットワーク構造は, モデルによって現れる典型的な構造を表している」というもう 1 つの仮定を設ける．すなわち, 現在のネットワーク構造がモデル (2) の定常期待状態であると仮定する．

ネットワークの平均的な構造が定常状態にあるとすると, (2) において, 期待値をとり, $t \to \infty$ としたものを $\phi^{(\infty)}(\cdot,\cdot) := \lim_{t\to\infty} E[\phi^{(t)}(\cdot,\cdot)]$ とすることで,

$$\phi^{(\infty)}(i,j) = \frac{\sum_{k\neq i,j} w_{kj}\phi^{(\infty)}(k,i) + w_{ki}\phi^{(\infty)}(k,j)}{\sum_{k\neq i,j} w_{kj} + w_{ki}}. \tag{3}$$

を得る．この方程式には全ての (i,j) ペアに対し $\phi^{(\infty)}(i,j) = 0$(あるいは $\phi^{(\infty)}(i,j) = 1$) という自明な解が存在するが, 次節でみるように, $\phi^{(\infty)}(i,j)$ を訓練データ E^L における実際のリンクラベルに固定することによって非自明な解を得ることができる．

4·2 パラメータ推定とリンクの予測

モデルのパラメータを推定し, $(i,j) \in E^U$ に対する $\phi^{(\infty)}(i,j)$ を予測するにあたり目的関数を定義する．リンクラベル $\phi(i,j)$ の対数尤度が

$$L_{ij} = \phi^{(\infty)}(i,j)\log\left(\frac{\sum_{k\neq i,j} w_{kj}\phi^{(\infty)}(k,i) + w_{ki}\phi^{(\infty)}(k,j)}{\sum_{k\neq i,j} w_{kj} + w_{ki}}\right)$$

$$+(1-\phi^{(\infty)}(i,j))\log\left(1-\frac{\sum_{k\neq i,j} w_{kj}\phi^{(\infty)}(k,i) + w_{ki}\phi^{(\infty)}(k,j)}{\sum_{k\neq i,j} w_{kj} + w_{ki}}\right)$$

のように書けることに注意すると, データ全体に対する対数尤度の和は以下のように書ける．

$$L(W) = \sum_{(i,j)\in E} L_{ij}. \tag{4}$$

我々はこれを目的関数として用いることにする[*1].

従って, 次の制約つき最適化問題を解けばよいことになる.

$$\begin{aligned}&\underset{W}{\text{maximize}}\ L(W)\\&\text{such that}\\&\quad\phi^{(\infty)}(i,j)=\frac{\sum_{k\neq i,j} w_{kj}\phi^{(\infty)}(k,i)+w_{ki}\phi^{(\infty)}(k,j)}{\sum_{k\neq i,j} w_{kj}+w_{ki}}\\&\qquad\text{for } (i,j)\in E^U,\\&\quad\text{and } \sum_{\ell,m} w_{\ell m}=1,\ w_{\ell m}\geq 0.\end{aligned}$$

ここで, 最大化すべき目的関数は E^L のリンクラベルの対数尤度の和であり, 制約はテストデータ E^U に対する定常性の条件である. テストデータがあらかじめ与えられているため, この問題はトランスダクションの問題であり, コピー確率 W だけでなく, 未知のリンクラベル $\{\phi^{(\infty)}(i,j)|(i,j)\in E^U\}$ も求めるべきパラメータとなる. この最適化問題を解くにあたり, W と $\{\phi^{(\infty)}(i,j)|(i,j)\in E^U\}$ を交互に最適化する反復型のトランスダクション法をとることにする.

まず, $\{\phi^{(\infty)}(i,j)|(i,j)\in E^U\}$ を固定して L を W について最大化するステップを考える. この最適化問題は非線形であり閉じた形での解を持たないため, 勾配に基づく数値的な最適化法を用いることにする. 目的関数の $w_{\ell m}$ についての勾配をとると次のようになる.

$$\frac{\partial L(W)}{\partial w_{\ell m}}=\sum_{i,j}\frac{\partial L_{ij}}{\partial w_{\ell m}}.$$

ここで

$$\begin{aligned}&\frac{\partial L_{ij}}{\partial w_{\ell m}}=\\&\phi^{(\infty)}(i,j)\cdot\left(\frac{\delta(m=j)\phi^{(\infty)}(\ell,i)+\delta(m=i)\phi^{(\infty)}(\ell,j)}{\sum_{k\neq i,j} w_{kj}\phi^{(\infty)}(k,i)+w_{ki}\phi^{(\infty)}(k,j)}\right.\\&\qquad\left.-\frac{1}{\sum_{k\neq i,j} w_{kj}+w_{ki}}\right)\\&+(1-\phi^{(\infty)}(i,j))\\&\cdot\left(\frac{\delta(m=j)(1-\phi^{(\infty)}(\ell,i))+\delta(m=i)(1-\phi^{(\infty)}(\ell,j))}{\sum_{k\neq i,j} w_{kj}(1-\phi^{(\infty)}(k,i))+w_{ki}(1-\phi^{(\infty)}(k,j))}\right.\\&\qquad\left.-\frac{1}{\sum_{k\neq i,j} w_{kj}+w_{ki}}\right)\end{aligned}$$

である.

W の全ての要素は 0 以上でかつ, 確率的な制約 (1) を満たすため, 我々は最適な W が疎であるときに収束が速いことが示されている指数勾配乗算 (exponentiated gradient) 型 [Helmbold 95] の最適化法を用いる. 繰り返しにおけるあるステップにおけるパラメータを W とすると, W から W' へ更新することで $L(W')$ が大きくなるようにしたい. 同時に, 新しい W' は W に十分に近いようにとるため, パラメータ間の距離のペナルティ $-d(W',W)$ を設ける. ここでは距離として以下の Kullback-Leibler(KL) ダイバージェンス

$$d(W',W)=\sum_{\ell,m} w'_{\ell m}\log\frac{w'_{\ell m}}{w_{\ell m}}$$

を用いることで, 指数勾配乗算型の更新則が導かれる. 各ステップにおいて, 新しいパラメータ W' を (1) の制約のもとで次の目的関数を最大化するように決める.

$$\eta L(W')-d(W',W),\ \eta>0,$$

ここで $\eta>0$ は 2 つの項のバランスをきめる定数である. ラグランジュ定数 γ を用いると, 最大化したい目的関数

$$F(W'):=\eta L(W')-d(W',W)-\gamma\Big(\sum_{\ell,m} w_{\ell m}-1\Big)$$

に対して, $L(W')$ を $L(W)$ を使って

$$L(W')=L(W)+\sum_{\ell,m}\frac{\partial L(W)}{\partial w_{\ell m}}(w'_{\ell m}-w_{\ell m}),$$

のように近似し, $w'_{\ell m}$ についての勾配を 0 とおくことで

$$\frac{\partial F(W')}{\partial w'_{\ell m}}=\eta\frac{\partial L(W)}{\partial w_{\ell m}}-\Big(\log\frac{w'_{\ell m}}{w_{\ell m}}+1\Big)+\gamma=0$$

となる. これを $\sum_{\ell,m} w_{\ell m}=1$ に注意して解けば, 次の指数勾配乗算型の更新則が導かれる.

$$w'_{\ell m}=Z^{-1}w_{\ell m}\exp\Big(\eta\frac{\partial L(W)}{\partial w_{\ell m}}\Big)\qquad(5)$$

ここで

$$Z:=\sum_{\ell,m} w_{\ell m}\exp\Big(\eta\frac{\partial L(W)}{\partial w_{\ell m}}\Big)$$

である.

次に, 未知変数は $\{\phi^{(\infty)}(i,j)|(i,j)\in E^U\}$ を求めるステップを考える. 前節でみたように, これらは定常性の条件 (3) を満たす必要があるため, この連立一次方程式を解いて未知変数の期待値を求めることになる.

上記を纏めると図 2 に示すアルゴリズムになる. [Step:3] は L を W について最大化するステップに, [Step:4] は未知変数を求めるステップに対応する.

4·3 逐次更新によるアルゴリズムの効率化

前節で示したアルゴリズムはバッチ型のため, 大きいサイズのデータの場合は効率が悪い. そこで, これを逐次更新によって近似したアルゴリズムを提案する. つま

*1 前節で述べたように, (4) は全てのリンクラベルの同時確率にはなっておらず, これを周辺確率 (2) の積として近似した平均場近似になっている.

Algorithm: Batch
[**Step:1**] $\ell \neq m$ である全ての (ℓ,m) について
$w_{\ell m} := \frac{1}{|V|-1}$ とする.
[**Step:2**] [Step:3] と [Step:4] を繰り返す.
[**Step:3**] (5) の更新式を用いて $L(W)$ を最大化する W を求める.
[**Step:4**] (3) を解いて, $(i,j) \in E^U$ について $\phi^{(\infty)}(i,j)$ を求める.

図 2 バッチ型トランスダクションアルゴリズム

Algorithm: Sequential
[**Step:1**] $\ell \neq m$ である全ての (ℓ,m) について
$w_{\ell m} := \frac{1}{|V|-1}$ とする.
[**Step:2**] (3) を解いて, $(i,j) \in E^U$ について $\phi^{(\infty)}(i,j)$ を求める.
[**Step:3**] [Step:4] を繰り返す.
[**Step:4**] (i,j) を一様にランダムに取り出す.
[**Step:4-a**] (6) によって $\{(\ell,m)|\ell \in \{1,2,\ldots,|V|\}, m \in \{i,j\}\}$ に対して $w_{\ell m}$ を更新する.
[**Step:4-b**] $(i,j) \in E^U$ ならば (7) によって $\phi^{(\infty)}(i,j)$ を更新する.

図 3 逐次型トランスダクションアルゴリズム

り, 図 2 の [Step:3] と [Step:4] を全ノードペアに対して繰り返すのではなく, 1 度に 1 つのノードペアを用いることにする.

(i,j) を現在注目しているノードペアとすると, [Step:3] では, (5) の代わりに, 訓練データを 1 つだけ使って, $\{(\ell,m)|\ell \in \{1,2,\ldots,|V|\}, m \in \{i,j\}\}$ に対して,

$$w'_{\ell m} = Z^{-1} w_{\ell m} \exp\left(\gamma \frac{\partial L_{ij}}{\partial w_{\ell m}}\right) \tag{6}$$

によって更新を行う. ここで, $\gamma > 0$ は更新の度合いを表す定数である. 尚, Z は全ての (ℓ,m) についての $w_{\ell m}$ の和であるが, Z の差分のみを計算することで, 1 回の更新は $O(|V|)$ で行うことができる.

同様に, [Step:4] では, (3) の連立 1 次方程式を解く代わりに, 冪乗法の 1 ステップを行うことにする. すなわち,

$$\phi^{(\infty)'}(i,j) := \frac{\sum_{k \neq i,j} w_{kj}\phi^{(\infty)}(k,i) + w_{ki}\phi^{(\infty)}(k,j)}{\sum_{k \neq i,j} w_{kj} + w_{ki}} \tag{7}$$

の更新を 1 回だけ行う.

この 2 つのステップを適当に繰り返すことで図 3 の逐次更新型のアルゴリズムが得られる.

5. 実　　験

この章では, 2 種類の生体ネットワークデータを用いて, 構造属性のみに基づくリンク予測問題における提案手法と既存の手法の性能を比較する.

5·1 既 存 手 法

まず, 提案手法と比較するために用いる既存のいくつかの指標 [Liben-Nowelly 04, Zan Huang 05] を説明する. これらの多くは対応するネットワーク構造の生成モデルから導かれたものである. ここで述べられる各指標は, 2 つのノードの間にリンクが存在する確からしさを示しており, ノードペアのランキングを与える. 従って, 適当な閾値で切る事でリンク予測を行うことができる.

すべての指標は, 存在するリンクのみを用いて定義されることに注意する. 以降, $\Gamma(i)$ をノード i の隣接ノードの集合 (リンクラベル 1 をもってノード i に連結したノードの集合) とする.

- Common neighbors [Newman 01]

$$\text{common} := |\Gamma(i)| \cap |\Gamma(j)|$$

Common neighbors 指標はノード i とノード j が共通の隣接ノードを多く持っているほど 2 つのノードの間にはリンクが現れやすいとする指標である.

- Jaccard 係数 [Baeza-Yates 99, Liben-Nowelly 04]

$$\text{Jaccard's} := \frac{|\Gamma(i)| \cap |\Gamma(j)|}{|\Gamma(i)| \cup |\Gamma(j)|}$$

Jaccard 係数は, 正規化された common neighbors 指標であり, 情報検索の分野で類似度として用いられる. 少数の隣接ノードをもつノードのリンクほど重宝される.

- Adamic/Adar [Adamic 03]

$$\text{Adamic/Adar} := \sum_{k \in |\Gamma(i)| \cap |\Gamma(j)|} \frac{1}{\log |\Gamma(k)|}$$

Adamic/Adar 指標は, Jaccard 係数と同じく正規化された common neighbors 指標であるが, Adamic/Adar 指標では, 隣接ノードごとに異なった重みを割り当てる. 少数の隣接ノードをもつノードに大きな重みが割り当てられる.

- Katz_β [Katz 53]

$$\text{Katz}_\beta := \sum_{l=1}^{\infty} \beta^l |\text{paths}_{i,j}^{(l)}|$$

Katz_β は common neighbor 指標の一般化であり, より遠くの関係を考慮する. $\text{paths}_{i,j}^{(l)}$ はノード i からノード j への長さ l のパスの数である. これは本質的にはカーネル法において 2 つのノード間の類似度を定義する拡散カーネル [Kondor 02] に等しい. 尚, 実験においては $\beta = 0.05$ とした.

- Preferential attachment [Newman 01, Barabási 02]

$$\text{preferential} := |\Gamma(i)| \cdot |\Gamma(j)|$$

Preferential attachment 指標は上記の指標とは若干異なり，「隣接ノードが多いノードほど新たなリンクを得られやすい」というスケールフリーネットワーク [Newman 03] の生成モデルに基づいた指標である．

5·2 実験の設定

実験には 2 種類の生体ネットワークを用いた．1 つは比較的小さなサイズの代謝ネットワーク，もう 1 つは中程度の大きさのタンパク質相互作用ネットワークである．

実験に用いた代謝ネットワークデータは，KEGG/PATHWAY データベース [Kanehisa 04] から Yamanishi ら [Yamanishi 05] によって抽出されたネットワークデータであり，S. Cerevisiae の代謝パスウェイを表している (図 1)．このネットワークでは，各タンパク質がノード，リンクが 2 つのタンパク質が連続する反応を触媒することを表す．ノード数は 618，リンク数が 2782 であり，全ノードペア数に占めるリンクの割合は 0.015 である．

また，タンパク質相互作用ネットワークデータの方は，von Mering ら [Mering 02] によって抽出されたネットワークであり，我々は Tsuda と Noble [Tsuda 04] による中程度の信頼度のネットワークを用いた．ノード数は 2617，リンク数が 11855 であり，全ノードペア数に占めるリンクの割合は 0.003 である．

どちらのデータにおいても，リンクの数は全ノードペア数に比較して極端に小さく，ネットワーク構造データとして典型的な，極めて偏りのあるデータであるといえる．実装においては，この偏りを補正するため図 3 の [Step:4] において，リンクを持つノードペアはリンクを持たないノードペアの割合での，またリンクをもたないノードペアはリンクを持つノードペアの割合での重み付きサンプリングを行った．

我々は全ノードペアの 66% を訓練データとして，残りをテストデータとして用い，3 分割の交差検定によって性能を測定した．前述したように両データとも偏りが大きいため，性能は precision-recall 曲線によって比較した．

5·3 実験結果

まず，各手法の予測精度を比較する．図 4 は代謝ネットワークデータに対する，また図 4 はタンパク質相互作用ネットワークデータに対する各手法の性能を，precision-recall 曲線で表したものである．表 1 はこれらの break-even 点 (precision=recall となる点) を纏めたものである．

全体的に提案手法がその他の手法を上回る予測性能を持っていることが確認できる．特に代謝ネットワークデータについてはその差が大きい．タンパク質相互作用ネットワークデータについては，特に break-even 点においてあまり著しい改善は見られないように見えるが，リンク予測においては，特に recall の低い領域 (左半分の領域) において，高い precision を持つことが重要である．なぜならば，実際のアプリケーションにおいてはリンク予測は非常に多くのノードペアの中から比較的少数の有望なノードペアを選び出してくるような使われ方 (例えば，実験的にまだ相互作用が確認されていないタンパク質のペアを発見するなど) をされるためである．このような視点から見ると提案手法はその他の手法に比べて非常に高い予測性能を持っているといえる．

次に，我々のトランスダクション型の定式化による効果を，トランスダクションを使わない場合 (単純に $(i,j) \in E^U$ に対して $\phi(i,j) = 0$ として，未知変数についての制約を用いずに学習を行う場合) と比較することで確認する．図 6 は，両者の代謝ネットワークデータにおける precision-recall 曲線を示したものである．これによりトランスダクションによるアプローチが予測性能を向上していることがわかる．

	metabolic	protein-protein
common	21.1%	37.9%
Jaccard's	30.2%	47.9%
Adamic/Adar	34.9%	50.3%
preferential	9.2%	25.4%
$\text{Katz}_{0.05}$	32.8%	27.6%
proposed	61.2%	52.6%

表 1 precision と recall の break-even 点による各手法の比較．

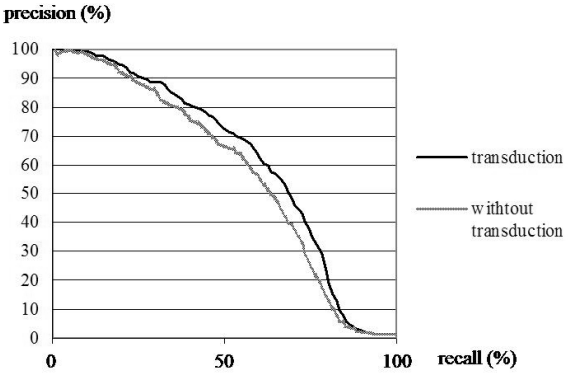


図 6 トランスダクションの効果を Precision-Recall 曲線で比較したもの．break-even 点はトランスダクションによる定式化の場合 61.2%，そうでない場合は 58.0%．

6. 関連研究

リンク予測は通常，ノードペアの分類問題あるいはランキング問題として定式化され，ノード属性と構造属性の 2 種類の情報が用いられる．構造属性の多くはネッ

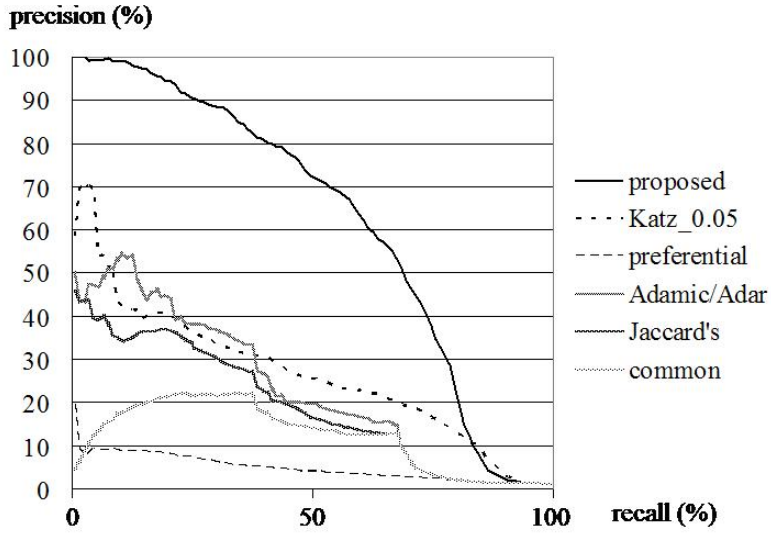


図 4 代謝ネットワークデータに対する Precision-Recall 曲線.

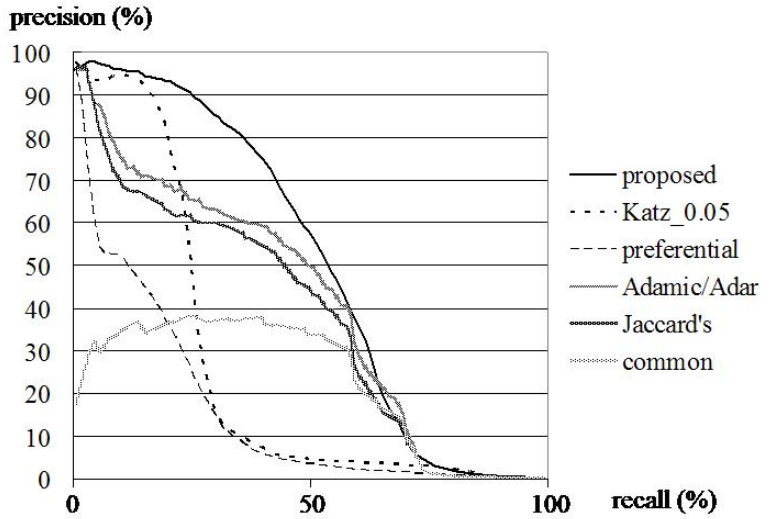


図 5 タンパク質相互作用ネットワークデータに対する Precision-Recall 曲線.

トワーク構造の生成モデル [Newman 01, Baeza-Yates 99, Liben-Nowelly 04, Adamic 03, Katz 53, Newman 01, Barabási 02] から導かれている場合が多く, 我々の提案したモデルも Kleinberg ら [Kumar 00, Kleinberg 99] によるノードコピーモデルのパラメトリックな場合としても解釈することができる. しかしながら我々の知る限りこれまでのところ, ノードコピーモデルに基づくリンク予測手法も, ネットワーク生成モデルに基づく学習可能なパラメータをもつリンク予測指標も提案されていない. また, 細かい違いを述べるなら, 我々のモデルではコピーされるのはリンクの存在ではなく, 存在／非存在を示すリンクラベルであるという点で Kleinberg らのモデルとは異なっている. 加えて, 彼らのモデルでは複数の枝が一度にコピーされうるのに対し, 我々のモデルでは一度に 1 つのリンクラベルのみがコピーされるが, これはコピー確率がパラメータであり必要に応じて調整されうるためである.

構造情報だけでなく, ノード情報も取り入れたリンク予測に教師あり学習を適用した試みとしては [Hasan 05] や [O'Madadhain 05] などがある. これらの研究では, 属性を統合する点に主眼が置かれており, 本論文のように個々の属性を議論するのとは視点が異なっている. 他にも, 確率的な関係学習のアプローチをリンク予測に適用した試みとして [Popescul 03] や [Taskar 03] などがあるが, これらの研究もやはり関係学習の一般的な枠組みを適用するというアプローチであり, 我々のリンク予測のために特化した個々の属性の性質に注目するという観点とは異なっている.

7. 結　　論

本論文では, ネットワーク構造におけるリンク予測の問題に対し, ノードの持つ情報によらず, 構造の持つ情報だけから予測を行う新しい手法を提案した. ネットワーク構造の時間的な変遷を, 枝の有無を表すリンクラベルがノードからノードへ確率的コピーされるとしてモデル化し, このモデルの定常状態をトランスダクションによる学習アルゴリズムによって求めることでリンク予測を行った. また, 生体ネットワークデータを用いた実験によって, 提案手法が良好な予測性能を持つことを確認した.

尚, 本論文ではリンクラベルがリンクの有無の 2 つのみであるとして議論を行ったが, 提案手法は, リンクラベルが 2 つ以上の場合にも容易に拡張することが可能である. 同様にしてリンクの向きを考慮することも可能であり, 今後は WWW の構造など, 有向リンクを持つ種々のネットワーク構造に対する適用も行っていく予定である.

最後に, 提案手法を実際に生体ネットワークの構造予測に用いるという観点から考察を行う. 実際的な観点からは, 本手法の直接的な適用に関しては, (1) 本手法で仮定しているネットワークのモデルが適切か, (2) 本論文での実験の設定 (枝が観測されない箇所ランダムに決まる) が, タンパク質の相互作用の予測問題の設定に合っているかどうか, という 2 点において未だ議論の余地があると思われる. (1) に関しては, [Newman 03] において, コピーに基づくモデルが生体ネットワークのモデルとしていくつか提案されている旨の指摘がある. 本論文での実験結果は, コピーに基づくモデルが生体ネットワークデータの性質に適しているという主張を部分的に支持していると思われるが, 今後はその他の種類のネットワーク構造データを用いた実験を行うことで提案手法の特徴や妥当性を調べていく必要がある. (2) に関しては, より現実的な問題設定では, あるタンパク質のセットについてはそれらの間の相互作用がある程度わかっているが, そのほかについては殆どわからないという場合が多く, このような場合には, 本手法が効果を発揮できない可能性が高いと思われる. 従って, なんらかの方法で [Yamanishi 05] などのノード情報に基づく手法や, 文献情報などのドメイン知識と組み合わせていくことが必要であると思われる. 生物学的データの観測の難しさに起因するデータの信頼度の問題への対処や, 「真に未知の」生物学的知識の発見のためにも, この方向は不可欠であろうと思われる.

謝　　辞

本研究を進めるにあたり貴重なご意見を頂いた京都大学化学研究所バイオインフォマティクスセンター　阿久津達也教授に感謝いたします。

◇ 参 考 文 献 ◇

[Adamic 03] Adamic, L. A. and Adar, E.: Friends and neighbors on the Web, *Social Networks*, Vol. 25, No. 2, pp. 211–230 (2003)

[Albert 99] Albert, R., Barabási, A., and Jeong, H.: Mean-field theory for scale-free random networks, *Physica A*, Vol. 272, pp. 173–187 (1999)

[Baeza-Yates 99] Baeza-Yates, R. A. and Ribeiro-Neto, B. A.: *Modern Information Retrieval*, ACM Press / Addison-Wesley (1999)

[Barabási 02] Barabási, A. L., Jeong, J., Néda., 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)

[Getoor 05] Getoor, L. and Diehl, C. P.: Link mining: a survey, *SIGKDD Explorations*, Vol. 7, No. 2, pp. 3–12 (2005)

[Hasan 05] Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M.: Link Prediction using Supervised Learning, in *Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)* (2005)

[Helmbold 95] Helmbold, D., Schapire, R., Singer, Y., and Warmuth, M.: A comparison of new and old algorithms for a mixture estimation problem, in *Proceedings of the Eighth Annual Workshop on Computational Learning Theory (COLT)*, pp. 69–78 (1995)

[Kanehisa 04] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M.: The KEGG resources for deciphering the genome, *Nucleic Acids Research*, Vol. 32, pp. D277–D280 (2004)

[Katz 53] Katz, L.: A new status index derived from sociometric analysis, *Psycometrika*, Vol. 18, No. 1, pp. 39–43 (1953)

[Kleinberg 99] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S.: The Web as a Graph: Measurements, Models and Methods, *Lecture Notes in Computer Science*, Vol. 1627, pp. 1–17 (1999)

[Kondor 02] Kondor, R. I. and Lafferty, J.: Diffusion Kernels on Graphs and Other Discrete Input Spaces, in *Proceedings of the Nineteenth International Conference on Machine Learning (ICML)*, pp. 315–322 (2002)

[Kumar 00] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E.: Stochastic Models for the Web Graph, in *IEEE Symposium on Foundations of Computer Science (FOCS)* (2000)

[Liben-Nowelly 04] Liben-Nowelly, 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)

[Mering 02] Mering, von C., Krause, R., Snel, B., Cornell, M., Olivier, S., Fields, S., and Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, Vol. 417, pp. 399–403 (2002)

[Newman 01] Newman, M. E. J.: Clustering and preferential attachment in growing networks, *Physical Review Letters E*, Vol. 64(025102), (2001)

[Newman 03] Newman, M. E. J.: The Structure and Function of Complex Networks, *SIAM Review*, Vol. 45, No. 2, pp. 167–256 (2003)

[O'Madadhain 05] O'Madadhain, J., Hutchins, J., and Smyth, P.: Prediction and ranking algorithms for event-based network data, *SIGKDD Explorations*, Vol. 7, No. 2, pp. 23–30 (2005)

[Popescul 03] Popescul, A. and Ungar, L. H.: Statistical relational learning for link prediction, in *IJCAI 2003 Workshop on Learning Statistical Models from Relational Data* (2003)

[Taskar 03] Taskar, B., Wong, M., Abbeel, P., and Koller, D.: Link prediction in relational data, in *Neural Information Processing System* (2003)

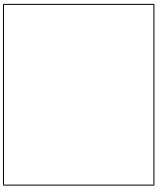
[Tsuda 04] Tsuda, K. and Noble, W. S.: Learning kernels

from biological networks by maximizing entropy, *Bioinformatics*, Vol. 20, No. Suppl. 1, pp. i326–i333 (2004)
[Wasserman 94] Wasserman, S. and Faust, K.: *Social Network Analysis*, Cambridge University Press (1994)
[Yamanishi 05] Yamanishi, Y., Vert, J.-P., and Kanehisa, M.: Supervised Enzyme Network Inference from the Integration of Genomic Data and Chemical Information, *Bioinformatics*, Vol. 21, pp. i468–i477 (2005)
[Zan Huang 05] Zan Huang, H. C., Xin Li: Link Prediction Approach to Collaborative Filtering, in *Proceedings of the Fifth ACM/IEEE-CS joint conference on Digital libraries (JCDL)* (2005)

〔担当委員：森下 真一〕

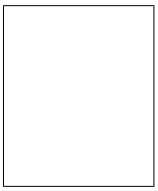
2006 年 8 月 15 日　受理

著　者　紹　介



鹿島　久嗣(正会員)

1999 年に京都大学工学研究科応用システム科学専攻にて修士課程修了．1999 年より，日本アイ・ビー・エム株式会社 東京基礎研究所に勤務．機械学習，データマイニング手法の開発と，バイオインフォマティクス，オートノミックコンピューティング，ビジネスインテリジェンス等への応用に従事．



安倍　直樹

1984 年に MIT にて修士課程修了．1989 年にペンシルバニア大学にてコンピュータサイエンスの Ph.D. を取得．1990 年から 2000 年まで NEC 中央研究所に勤務．1998 年から 2000 年まで東京工業大学にて客員助教授を務める．2001 年より米国 IBM T.J. ワトソン研究所に勤務．機械学習，データマイニング手法の開発と，ビジネスの分析や最適化への応用に従事．