リンク伝播法：リンク予測のための半教師付き学習法

Link Propagation: A Semi-supervised Approach to Link Prediction

鹿島 久嗣 [1*]　加藤 毅 [2]　山西 芳裕 [3]　杉山 将 [4]　津田 宏治 [5]
Hisashi Kashima[1] Tsuyoshi Kato[2] Yoshihiro Yamanishi[3] Masashi Sugiyama[4] Koji Tsuda[5]

[1] IBM 東京基礎研究所　IBM Tokyo Research Laboratory
[2] お茶の水女子大学　Ochanomizu University
[2] パリ国立高等鉱業学校　Mines ParisTech
[4] 東京工業大学　Tokyo Institute of Technology
[5] マックスプランク研究所　Max Planck Insitute for Biological Cybernetics

Abstract: We propose Link Propagation as a new semi-supervised learning method for link prediction problems, where the task is to predict unknown parts of the network structure by using auxiliary information such as node similarities. Since the proposed method can fill in missing parts of tensors, it is applicable to multi-relational domains, allowing us to handle multiple types of links simultaneously. We also give a novel efficient algorithm for Link Propagation based on an accelerated conjugate gradient method.

1　概要

1.1　リンク予測問題

世の中には、単に、その事象の構成要素によってだけではなく、それらの間の（静的あるいは動的な）**「関係」によって記述されうる事象が多く存在する**。例えば、

- 生体ネットワーク： タンパク質と、それらの間の相互作用関係
- ソーシャルネットワーク： 人間やグループと、それらの間の交友／所属の関係
- オンライン・ショッピング： ユーザーとオンライン広告や商品、それらの間の（「クリックする」「買う」などといった）関係

などが例として挙げられる。

リンク予測問題は、関係の予測を目的とした問題である。これは、従来の予測問題が、主に構成要素自身の性質を予測することを目的としていたのに比べて、より一般的な予測問題であるといえる。一般に、リンク予測の問題は、ネットワーク構造の観測されている部分から、観測されていない部分 (あるいは、将来のネットワークの構造) を予測する問題として定式化され、機械学習でいうところの教師付きの予測問題として捉えることができる。データマイニングにおいては、リンクマイニングと呼ばれるテーマの一タスクとして認識されている [4]。

上で挙げたような、生体ネットワークやソーシャルネットワークの分析、オンライン・マーケティングなどの応用において、**リンク予測は基本的かつ重要な技術である**。例えば、タンパク質の相互作用ネットワークにおいて、既知の相互作用をもとに、未知の相互作用を予測し、優先付けすることができれば、実験のコストを軽減することができる。あるいは、ソーシャルネットワーク・サービスにおいて、既存の人間関係や記事の購読、グループなどへの所属をもとにして、友人や記事、グループなどの推薦を適切に行うことができれば、滞在時間の延長に結びつけることができる。また、オンライン・マーケティングにおいては、過去の実績データをもとに、ユーザーが、どのオンライン広告をクリックするか、あるいは、どの商品を購入するか、などを予測することができれば、売り上げを向上することが期待できる。

*連絡先： 日本アイ・ビー・エム株式会社 東京基礎研究所
〒 242-8502 神奈川県大和市下鶴間 1623-14
E-mail: hkashima@jp.ibm.com

1.2　リンク予測問題へのアプローチ

リンク予測問題への機械学習アプローチは、大きく分けて、次の 2 つに分類することができる。

1) ペアワイズ予測モデル 任意の 2 構成要素間のリンクの有無の予測 【本論文で採用】

2) 関係ネットワークモデル ネットワーク構造全体の生成モデルによる予測

前者のペアワイズ予測モデルは、2 つの構成要素を入力とした予測を行うという点で、通常の（単体の構成要素を入力とした予測を行う）教師付き予測の枠組みに帰着できるため、比較的大きな問題が扱いやすいという利点がある。ペアワイズ予測モデルでは、各リンクの有無は独立である、という単純化を行っているが、リンクの間の依存関係もモデルに取り入れようとすると、後者の関係ネットワークモデル、すなわち、より一般的な、ネットワーク構造全体の生成を考えるモデ

ルが必要になる。しかしながら、リンクの間の依存関係が、推論を複雑にするため、スケーラビリティは限定されてしまう（通常、NP 困難となる）。

本論文では、スケーラビリティの観点から、**(1) ペアワイズ予測モデルを採用する**ことにする。

さて、モデルとして、ペアワイズ予測モデルを採用するとした場合に、推論に用いる基本的な考え方は、さらに 2 種類に分類することができる。

1-a) ノード特徴ベース推論 類似した構成要素同士の間にリンクが張られるとする

1-b) ペアワイズ特徴ベース推論 類似した「構成要素のペア」同士は、リンクの有無も似ているとする【本論文で採用】

前者の仮定（ノード特徴ベース推論）は、各ノードの持つ特徴をベースに、よく似たノード同士の間にはリンクが張られるであろうという、「似たもの同士のネットワーク」を仮定しているといえる。一方、後者の仮定（ペアワイズ特徴ベース推論）は、構成要素のペアが 2 つあるとき、これらが（ペアとして）類似しているならば、リンクの有無に関しても類似している、という、一段複雑な仮定を行っている。例えば、2 つの構成要素が、それぞれある特徴をもっているときにリンクが存在する（あるいは、存在しない）という場合がこれにあたり、タンパク質の相互作用において、片方のタンパク質が、ある構造 A を持ち、もう片方のタンパク質が、別の構造 B を持つときに相互作用がある、といったことを考慮することができる。また、ノード特徴ベース推論が、構成要素の均一性を暗に仮定しているのに対し、ペアワイズ特徴ベース推論は、異種混合的な状況においても有効であるという特徴がある。

本論文では、適用範囲の広さから、後者の **(1-b) ペアワイズ特徴ベース推論を採用する**ことにする。前者がノードレベル（$O(N)$ のレベル）での推論を行うのに対し、後者はリンクレベル（$O(N^2)$ のレベル）での推論を行うため、必然的にスケーラビリティの問題が発生する。この問題を解決するのが、本論文の狙いでもある。

1.3 提案手法の特長と技術的なチャレンジ

本論文で提案するリンク予測手法の特長として、以下の 3 点を挙げることができる。

半教師付き予測 リンク予測問題においては、リンク未知の構成要素ペアが予め分かっている場合がしばしばあるが、これら、リンク未知の構成要素ペアの情報も用いた予測を行うことができる

複数種リンクの同時予測 複数のリンクの種類（例えば、オンライン・ショッピングにおける「クリック」「購買」「評価」など）がある場合に、これらの間の相関を利用して、同時に予測を行うことができる

ノード情報とリンク情報の同時利用 ノードの持つ情報（例えば、タンパク質相互作用における、タンパク質の配列や、遺伝子発現強度など）と、リンクの持つ構造情報の両方を、ノードの類似度という形で統一的に用いることができる

リンク未知の構成要素ペアの情報を有効活用するため半教師付き予測であることと、ペアワイズ特徴ベース推論を採用することから、スケーラビリティの問題が発生する。この**スケーラビリティの問題を軽減することが、本論文における技術的なチャレンジ**となる。

1.4 本論文の貢献

本論文の貢献をまとめると、以下の 4 点である。

半教師付き予測 ラベル伝播法をベースにして、初めての半教師付きリンク予測法を提案する

複数種リンクの同時予測 初めての、複数種類のリンクを扱うことのできるリンク予測法を提案する

新しいペアワイズ類似度 ノード類似度のクロネッカー和による、新しいペアワイズ類似度を提案する

効率的な予測アルゴリズム 共役勾配法を加速することによって、効率的な予測アルゴリズムを提案する

なお、**より詳細な導出や、追加実験の結果、従来研究の調査、今後の方向性などについては、原論文 [8] を参照されたい**。

2 複数種リンク予測問題の定式化

リンク予測問題は、通常、任意の 2 つの構成要素（以下、ノードと呼ぶことにする）の間のリンクの強さを予測する問題として捉えることができる。特に、本論文では、複数種類のリンクがあるようなリンク予測問題を考える。2 つのノード集合 $X \equiv \{x_1, x_2, \ldots, x_M\}$ と $Y \equiv \{y_1, y_2, \ldots, y_N\}$、また、リンクの種類 $Z \equiv \{z_1, z_2, \ldots, z_T\}$ があるとする（X と Y は同一であってもよいし、別々の集合であってもよい）。また、それぞれの集合の大きさを $M \equiv |X|, N \equiv |Y|, T \equiv |Z|$ とする。$T = 1$ のとき、通常扱われる、単一種のリンク予測問題となる。

目的は、任意の 3 つ組 $(x_i \in X, y_j \in Y, z_k \in Z)$ に対して、そのリンクの強さを予測することとなる。従って、予測を $M \times N \times T$ の大きさを持つ 3 階のテンソル $\mathcal{F}$ として表わし、これをリンク強度と呼ぶことにする。その (i, j, k) 成分 $[\mathcal{F}]_{i,j,k}$ は、X の i 番目のノードと Y の j 番目のノードの間に、Z の k 番目の種類のリンクが存在する確からしさを表す。

ここで、ネットワーク構造の既知部分についての情報として、別の $M \times N \times T$ の大きさを持つ 3 階テンソル $\mathcal{F}^*$ を用意する。$\mathcal{F}^*$ の各要素は以下のように定義される。

$$[\mathcal{F}^*]_{i,j,k} \equiv \begin{cases} +\epsilon^+ & 3 \text{ つ組 } (x_i, y_j, z_k) \text{ にリンクが存在する場合} \\ -\epsilon^- & 3 \text{ つ組 } (x_i, y_j, z_k) \text{ にリンクが存在しない場合} \\ 0 & \text{リンク未知の場合} \end{cases}$$

ϵ^+ と ϵ^- は適当な正の定数であり、回帰とフィッシャー判別との関係 [2] に倣い、それぞれリンクの存在比の逆数、リンクの非存在比の逆数にとることにする。

さらに我々は、補助情報として、ノードやリンクの持つ情報を、対称で非負の類似度行列の形で得られるものとする。X と Y に含まれるノード間の類似度行列を、それぞれ W_X、W_Y とする。例えばタンパク質の場合には、そのアミノ酸配列や、遺伝子発現強度などのノードの持つ情報の類似度であったり、隣接行列そのものや、共通隣接ノード数によって作られた類似度などの、構造的な類似度を用いることができる。また、Z に含まれるリンク種間の類似度行列を W_Z とする。これは、2 つのリンク種間の共起度合いや、事前知識によって設計する。なお、これらの類似度行列は、対称で非負であるとする（半正定である必要はない）。

まとめると、複数種リンク予測問題の入出力は以下のようになる。

入力:
- 3 つの対称で非負な類似度行列 W_X、W_Y、および W_Z
- 既知のネットワーク構造を表す 3 階テンソル $\mathcal{F}^*$

出力: ネットワーク構造の未知部分に対するリンク強度の予測値を表す 3 階テンソル $\mathcal{F}$

3 提案手法

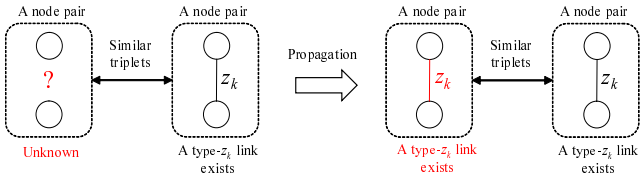
3.1 リンク伝播法

我々の問題設定は、リンク未知のペアについても、ノードの持つ情報が類似度行列の形で与えられているため、半教師付きの予測問題（より厳密にはトランスダクション）として捉えるのが妥当である。そこで、半教師付き予測の代表的手法であるラベル伝播法 [14, 15] を用いることにする。ラベル伝播は、元来、ノードの性質を予測するための手法であり、ラベルなしノードの作る構造を利用するために、ラベル伝播原理「類似したノード同士は、同じラベル（性質）を持つ可能性が高い」という仮定をもとに推論を行う。

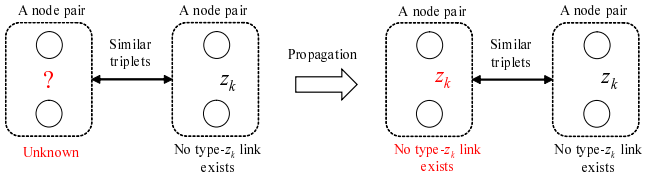
我々は、リンク予測問題が、ある 2 つのノードと、あるリンク種の 3 つ組みに対し、「リンクあり」または「リンクなし」のラベルを予測する問題と解釈できることから、**ラベル伝播法を (ノード, ノード, リンク種) の 3 つ組に対して適用する**ことにする。そこで、ラベル伝播が用いる仮定を 3 つ組みに対して拡張した、リンク伝播原理「類似したノードペア同士は、同じリンクを持つ可能性が高い」という仮定を用いる。

リンク伝播原理に基づく推論を、最適化問題の（最小化すべき）目的関数の形で表すことにする。

$$J(\mathcal{F}) \equiv \frac{\sigma}{2} \sum_{i,j,k,\ell,m,n} [\mathsf{W}]_{ijk,\ell mn} \left([\mathcal{F}]_{ijk} - [\mathcal{F}]_{\ell mn}\right)^2 + \frac{1}{2} \sum_{(i,j,k)\in E} \left([\mathcal{F}]_{ijk} - [\mathcal{F}^*]_{ijk}\right)^2 \quad (1)$$



(a) リンクの存在が伝播する例



(b) リンクの不在が伝播する例

図 1: リンク伝播原理： (a) リンクの存在が伝播する場合と (b) リンクの不在が伝播する場合。2 つの 3 つ組が類似しているならば、それらに対するリンクの有無が一致する可能性も高い

ここで、1 項目がリンク伝播原理を表しており、W は 2 つの 3 つ組間の類似度を表す行列（後ほど定義する）とする。3 つ組の対 (x_i, y_j, z_k) と (x_ℓ, y_m, z_n) に対する $[\mathsf{W}]_{ijk,\ell mn}$ が大きいほど、これらに対するリンク強度の予測値が近くなることが要求される。一方、2 項目は、リンクが存在する部分については予測値を正の値 ϵ^+ に、リンクが存在しない部分については予測値を負の値 $-\epsilon^-$ に、未知の部分については 0 に近づけるように働く。これら 2 つの項を、正の定数 σ でバランスをとり、最小化する $\mathcal{F}$ をもって予測とする。

式 (1) をテンソルを用いて書き直すと、

$$J(\mathcal{F}) = \frac{\sigma}{2} \mathbf{vec}\,(\mathcal{F})^\top \mathsf{L}\,\mathbf{vec}\,(\mathcal{F}) + \frac{1}{2} \parallel \mathbf{vec}\,(\mathcal{F}) - \mathbf{vec}\,(\mathcal{F}^*) \parallel_2^2 \quad (2)$$

と書ける。ここで、L は、$\mathsf{L} \equiv \mathsf{D} - \mathsf{W}$, によって定義されるラプラシアン行列である。なお、D は、W の各行の和を対角成分に持つような、対角行列とする。また、$\mathbf{vec}\,(\mathcal{F})$ は、テンソル $\mathcal{F}$ の列ベクトル（mode-1 fibers）を、積んで構成されるベクトルを表す。

式 (2) を $\mathbf{vec}\,(\mathcal{F})$ で微分して $\mathbf{0}$ と置くと、結局、以下の**巨大な連立方程式を解くことで予測が求まる**ことがわかる。

$$(\sigma \mathsf{L} + \mathsf{I}_{MNT})\,\mathbf{vec}\,(\mathcal{F}) = \mathbf{vec}\,(\mathcal{F}^*) \quad (3)$$

3.2 3 つ組間の類似度行列の設計

式 (1) において与えられていなかった、3 つ組間の類似度行列 W を、与えられた 3 つの（要素間、リンク間）類似度行列 W_X、W_Y、W_Z から構成する。本論文では、**3 つ組間の類似度行列 W の構成方法として、ク**

ロネッカー 積 類似度とクロネッカー 和 類似度の 2 種類を考える。特に、クロネッカー和類似度は、本論文において初めて用いられる類似度定義である。

まず、クロネッカー積類似度を定義する。2 つの 3 つ組に対する類似度の定義として、3 つ組のそれぞれの対応する要素が全て類似している場合に、3 つ組同士が類似している、と定義する（図 2(a)）のは自然である。これは、3 つの類似度行列のクロネッカー積として、

$$\mathsf{W} \equiv \mathsf{W}_Z \otimes \mathsf{W}_Y \otimes \mathsf{W}_X \tag{4}$$

のように書くことができる。これはちょうど、ペアワイズカーネル [1, 10] を 3 つ組に拡張した形になっている。3 つの類似度行列が半正定であるとき、クロネッカー積類似度は、3 つの類似度行列が作る特徴空間の積空間における内積（カーネル）となる。従って、クロネッカー積類似度の作る特徴空間は、時に、複雑すぎることが懸念される。そこで、もう 1 つの類似度定義として、類似する 3 つ組ペアを大きく限定する、クロネッカー和類似度

$$\begin{aligned}\mathsf{W} &\equiv \mathsf{W}_Z \oplus \mathsf{W}_Y \oplus \mathsf{W}_X \qquad (5)\\ &= \mathsf{W}_Z \otimes \mathsf{I}_N \otimes \mathsf{I}_M + \mathsf{I}_T \otimes \mathsf{W}_Y \otimes \mathsf{I}_M + \mathsf{I}_T \otimes \mathsf{I}_N \otimes \mathsf{W}_X\end{aligned}$$

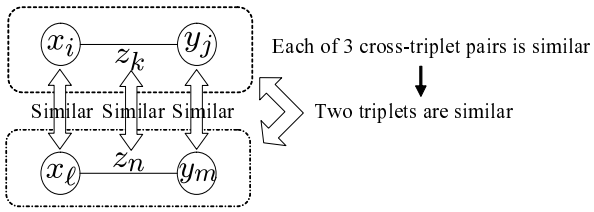
を定義する。これは、3 つ組のそれぞれの対応する要素のうち 2 つが同一で、且つ、残りの 1 組が類似しているときに、3 つ組同士が類似している、と定義する（図 2(b)）。

定義から分かるように、クロネッカー積類似度は任意の 3 つ組みペアに対して、0 より大きい類似度を取りうるのに対して、クロネッカー和類似度のほうは、遥かに少ない 3 つ組みペアのみが 0 より大きい類似度をとりうる。一見、これは類似度情報を十分に生かしきれていないようにも思われるが、後の実験結果で示されるように、クロネッカー和類似度は半教師付き予測手法であるリンク伝播法との相性がよい。これはおそらく、定義上類似度が 0 になってしまう 3 つ組みペアが、0 以上の類似度をもつ 3 つ組みペアを通じて比較できるようになるためと思われる。

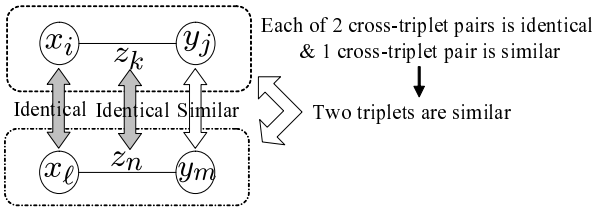
3.3 効率的な予測アルゴリズム

我々の目的は式 (4) もしくは式 (5) によって定義される 3 つ組み間の類似度行列をもった連立方程式 (3) を解くことである。しかしながら、一見してもわかるように、類似度行列のサイズが $MNT \times MNT$ と非常に大きくなってしまうため、これをメモリ内に明示的に構築して、連立方程式を解くのは適切ではない。

我々は、線型方程式の標準解法の 1 つである**共役勾配法をベースに効率的な解法を構成する**。共役勾配法のアルゴリズム [5] に、我々の連立方程式 (3) を当てはめたものが、アルゴリズム 1 である（但し、テンソルを用いた表記になっていることに注意する）。2 行目と 4 行目に現れる $\mathcal{L}^{\{\text{PROD}|\text{SUM}\}}$ は、クロネッカー積類似度を用いる場合には以下の $\mathcal{L}^{\text{PROD}}$ で、クロネッカー和類似度



(a) クロネッカー積類似度



(b) クロネッカー和類似度

図 2: 3 つ組同士の類似度定義： (a) クロネッカー積類似度 (b) クロネッカー和類似度

を用いる場合には以下の $\mathcal{L}^{\text{SUM}}$ によって置き換える。

$$\begin{aligned}\mathcal{L}^{\text{PROD}}(\mathcal{F}) &\equiv (\mathsf{D}_Z \otimes \mathsf{D}_Y \otimes \mathsf{D}_X \\ &\quad -\mathsf{W}_Z \otimes \mathsf{W}_Y \otimes \mathsf{W}_X)\,\mathbf{vec}\,(\mathcal{F}) \qquad (6)\\ \mathcal{L}^{\text{SUM}}(\mathcal{F}) &\equiv (\mathsf{D}_Z \oplus \mathsf{D}_Y \oplus \mathsf{D}_X \\ &\quad -\mathsf{W}_Z \oplus \mathsf{W}_Y \oplus \mathsf{W}_X)\,\mathbf{vec}\,(\mathcal{F}) \qquad (7)\end{aligned}$$

この導出には、若干の計算 [8] が必要だが、ここでは省略する。

ここで計算上のボトルネックとなるのが、式 (6) と式 (7) の評価であり、類似度行列のサイズの問題がここに凝縮された形となる。この 2 つの式の評価を高速化することが、アルゴリズムを現実的なものにするかどうかの分かれ目となる。更に言えば、ペアワイズ特徴ベース推論（1.2 節参照）を行うほぼ全ての手法は、これに類する問題を抱えており、現在までのところ解決されていない。

実は、以下のように、**テンソルのモード積を用いることで、計算を大幅に高速化することができる**。

$$\begin{aligned}\mathcal{L}^{\text{PROD}}(\mathcal{F}) &= \mathcal{F} \times_1 \mathsf{D}_X \times_2 \mathsf{D}_Y \times_3 \mathsf{D}_Z \\ &\quad - \mathcal{F} \times_1 \mathsf{W}_X \times_2 \mathsf{W}_Y \times_3 \mathsf{W}_Z \qquad (8)\\ \mathcal{L}^{\text{SUM}}(\mathcal{F}) &= \mathcal{F} \times_1 (\mathsf{D}_X - \mathsf{W}_X) + \mathcal{F} \times_2 (\mathsf{D}_Y - \mathsf{W}_Y) \\ &\quad + \mathcal{F} \times_3 (\mathsf{D}_Z - \mathsf{W}_Z) \qquad (9)\end{aligned}$$

$\times_1$ などの演算は、テンソルのモード積であり、（詳細な定義は [9] などに譲るが）テンソルの各モードのベクトルに対して、行列を掛ける演算である。式 (6) と式 (7) が、大きさ $MNT \times MNT$ の巨大な類似度行列と、大きさ MNT のベクトルの掛け算を行うのに対し、式

Algorithm 1 リンク伝播法（入力は $\mathcal{F}^*, \mathcal{G}, \mathsf{W}_X, \mathsf{W}_Y, \mathsf{W}_Z, \sigma, \epsilon$）$\mathcal{L}^{\{\text{PROD}|\text{SUM}\}}$ は、$\mathcal{L}^{\text{PROD}}$（式 (8)）あるいは $\mathcal{L}^{\text{SUM}}$（式 (9)）で置き換える

$\mathcal{F}(0) := \mathcal{F}^*$
$\mathcal{R}(0) := -\sigma\mathcal{L}^{\{\text{PROD}|\text{SUM}\}}(\mathcal{F}(0))$, and $\mathcal{P}(0) := \mathcal{R}(0)$
for $t = 0, 1, 2, \ldots$ **do**
  $\mathcal{Q}(t) := \sigma\mathcal{L}^{\{\text{PROD}|\text{SUM}\}}(\mathcal{P}(t)) + \mathcal{P}(t)$
  $\alpha(t) := \dfrac{\left\langle \mathcal{R}(t), \mathcal{P}(t) \right\rangle}{\left\langle \mathcal{P}(t), \mathcal{Q}(t) \right\rangle}$
  $\mathcal{F}(t+1) := \mathcal{F}(t) + \alpha(t)\mathcal{P}(t)$
  $\mathcal{R}(t+1) := \mathcal{R}(t) - \alpha(t)\mathcal{Q}(t)$
  $\beta(t) := \frac{\|\mathcal{R}(t+1)\|_2^2}{\|\mathcal{R}(t)\|_2^2}$
  if $\frac{\|\mathcal{R}(t+1)\|_2}{\|\mathcal{R}(0)\|_2} < \epsilon$, **return** $\mathcal{F}(t+1)$
  $\mathcal{P}(t+1) := \mathcal{R}(t+1) + \beta(t)\mathcal{P}(t)$
end for

(8) と式 (9) では、ノード間あるいはリンク間の類似度行列（$M \times M$ などの大きさ）と、行列（$M \times NT$ などの大きさ）の掛け算となるため、大幅な速度向上を行うことができる。

4 実験

4.1 実験の設定

単一種のリンク予測問題と、複数種類のリンク予測実験の、2 種類の実験を行う。

単一種のリンク予測問題では、既存のペアワイズ特徴ベース推論であるペアワイズ・カーネル法 SVM と比較して、**提案手法が、遥かに高速で、高い精度の予測を行う**ことを示す。

複数種類のリンク予測実験では、単一種のリンク予測問題を別々に行うよりも**複数種類のリンクを同時に予測することで、高い精度での予測が行える**ことを示す。

なお、他のデータや、3 つのネットワークの同時予測による結果などは、原論文 [8] を参照されたい。

4.2 単一種類のリンク予測実験

まず、単一種類のリンク予測実験において、提案手法とペアワイズ・カーネル法 [1, 10] との比較を行う。

ペアワイズ・カーネル法で用いられるカーネル行列は、クロネッカー積によって構成されるため、これを明示的に作ることはできない。従って、オンライン型のアルゴリズムである passive-aggressive アルゴリズム（PA-I タイプ）[3] を用いた。正則化パラメータは $C = 1$ を用い、全データを 3 週分、処理を行った。また、通常のクロネッカー積によるカーネル行列に加え、クロネッカー和によるカーネル行列を用いた実験も行った。提案手法については、$\sigma = 0.001$ とした。

ベンチマークデータとしては、Yamanishi ら [13] に倣い、酵母菌内の代謝ネットワークデータ [7] を用いた。このネットワークは、618 のノード（タンパク質）と 2782 のリンクを持つ。リンクは、2 つのタンパク質が連続した反応を触媒することを意味する。類似度行列は、遺伝子発現、局在部位、系統発生の類似度から作られる 3 種類を用いた[1]。

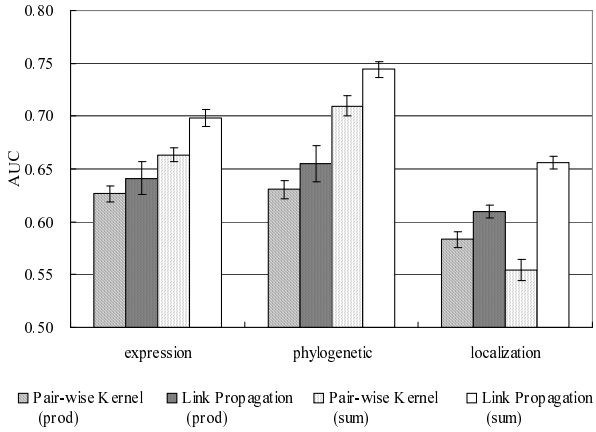


図 3: 単一種類のリンク予測の精度比較

全ノード対のうち、ランダムに選んだ 10% についてのみ、リンクの有無が分かっているものとし、残りのものに対する AUC 値を計測した。これを 10 回行い、AUC 値の平均と標準偏差によって評価を行った。図 3 が、3 種類の類似度行列（「expression」「phylogenetic」「localization」）それぞれを用いた、提案手法（「Link Propagation」）とペアワイズ・カーネル法（「Pair-wise Kernel」）による AUC の平均と標準偏差を示したものである。「(prod)」と「(sum)」は、それぞれ、クロネッカー積類似度（カーネル）とクロネッカー和類似度（カーネル）による結果を指す。

結果から、提案手法の予測精度は、ペアワイズ・カーネル法の結果を上回っていることが分かる。また、クロネッカー和類似度が、クロネッカー積類似度より予測精度が高いのは、クロネッカー和類似度が、半教師付き予測法との相性がよいことを示していると思われる。

次に、計算時間の比較を行う。図 4 は、代謝ネットワークデータ（618 ノード）に加え、タンパク質相互作用ネットワーク（2617 ノード）[12] と、論文の共著ネットワーク[2]（2865 ノード）を用いた、学習と予測をあわせた計算時間を比較したものである。ペアワイズ・カーネル法は、全データを一周分しか用いていないが、それにも関わらず、提案手法の方が遥かに高速であることがわかる。

4.3 複数種類（2 種類）のリンク予測実験

次に、複数種類のリンクを同時に予測することで、リンク種ごとに別々に予測を行うよりも、予測精度が良くなることを示す。

[1] http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/ismb05/ より入手可能
[2] http://ai.stanford.edu/~gal/data.html より入手可能

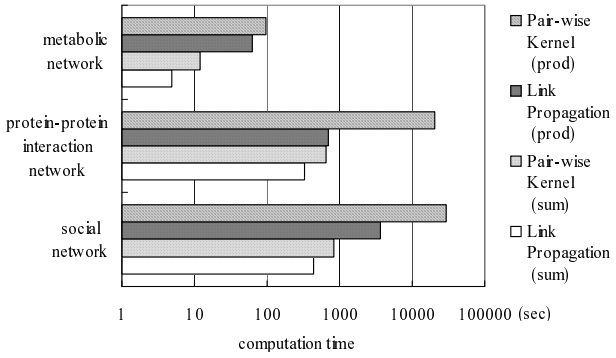
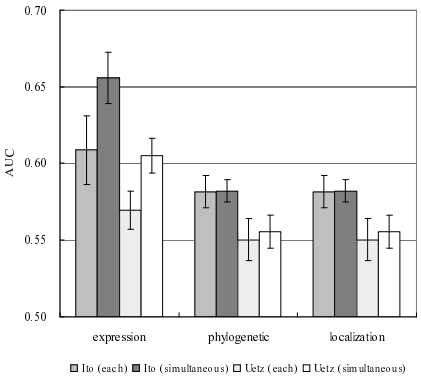


図 4: 単一種類のリンク予測における実行時間比較

ここでは、2 つの研究室で得られた、酵母のタンパク質相互作用ネットワークデータ [6, 11] を用いる。2 つの研究室それぞれで特定された相互作用を、2 種類のリンクとみることで、同時に予測することを行う。ネットワークは 1422 のノードを持ち、リンクはそれぞれ 744 と 888 あり、このうち重複は 123 であった。

類似度行列は、やはり Yamanishi ら [13] と同様の方法で 3 種類（expression、localization、phylogenetic）を構成した。リンク間類似度は、単純に 1 とした。今回は、ある程度のリンクの重なりが必要であるため全ノード対のうち、ランダムに選んだ 50% について、リンクの有無が分かっているものとした。

図 5 は、3 種類の類似度行列（「expression」「phylogenetic」「localization」）それぞれを用いた、2 つのネットワークの同時予測（「Simultaneous」）と個別予測（「Individual」）による AUC の平均と標準偏差を示したものである。「(Ito)」と「(Uetz)」は、2 つの研究室のネットワークのそれぞれに対する結果を示す。また、上がクロネッカー積類似度（カーネル）、下がクロネッカー和類似度（カーネル）による結果を指す。結果より、同時予測によって、精度が向上する場合が多いことが分かる。また、この実験においても、クロネッカー和類似度の方が性能が良いことが分かる。



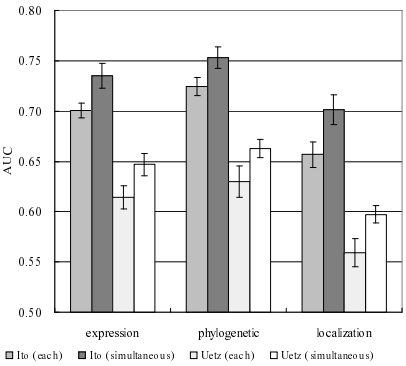


図 5: 2 種類のリンク予測の精度比較。上がクロネッカー積類似度（カーネル）、下がクロネッカー和類似度（カーネル）による結果

参考文献

[1] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(Suppl. 1):i38–i46, 2005.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

[4] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explorations*, 7(2):3–12, 2005.

[5] G. H. Golub and C. F. V. Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, 1996.

[6] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interacome. *Proc. the National Academy of Sciences*, 98(8):4569–4574, 2001.

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

[8] H. Kashima, T. Kato, Y. Yamanishi, M. Sugiyama, and K. Tsuda. Link propagation: A fast semi-supervised learning algorithm for link prediction. In *SDM*, 2009.

[9] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. Technical Report SAND2007-6702, Sandia National Laboratories, 2007.

[10] S. Oyama and C. D. Manning. Using feature conjunctions across examples for learning pairwise classifiers. In *ECML*, pages 322–333, 2004.

[11] P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, and et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

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

[13] Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21:i468–i477, 2005.

[14] D. Zhou, O. Bousquet, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS 16*, 2004.

[15] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.