解　説

ネットワーク構造予測

Survey of Network Structure Prediction Methods

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

keywords: link prediction, link mining, biological network analysis, social network analysis

Summary

Recently, there has been a surge of interest in the study of analytical methods for network structured data such as WWW, social networks, and biological networks, and *link mining* has become a popular subarea of data mining. In this survey, we focus on the link prediction problem, which is the task of predicting unobserved portion of the network structure, i.e. hidden links, from the observed part of the network structure (or to predict the future structure of the network given the current structure of the network.) Link prediction has several important applications including predicting relations among participants such as friendships, communications and collaborations in social networks. In the field of bioinformatics, predicting protein-protein interactions and regulatory relationships can provide guidance on the design of experiments for discovering new biological facts. We introduce several approaches for the link prdiction problem, including methods based on various information about network structures, and discuss relation with other problems.

1. は　じ　め　に

近年, WWW のリンク構造はもとより, SNS（ソーシャルネットワークサービス）の人気や, 生物学における生体ネットワークの重要性の認識の高まりとともに, ネットワーク構造をもったデータの解析が注目を集めている. 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].

データマイニングや機械学習の分野でも, ネットワーク構造解析への興味は高まりを見せており, 「リンクマイニング」などとよばれ, その重要性が広く認識されつつあるとともに, 現在も非常に研究が盛んなトピックである. 本稿で解説するリンク予測の問題は, リンクマイニングで扱うタスクのひとつであり, ネットワーク構造の観測される部分を手がかりに, 残りの部分の構造を予測する（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. その応用は, 例えば, 社会ネットワークにおける人間関係や将来の相互作用を予測したり, バイオインフォマティクスにおいては, タンパク質の相互作用や遺伝子間の制御関係の予測を行ったりなど多岐にわたる.

本稿では, リンク指標やネットワーク構造の確率モデルなどリンク予測問題に対するいくつかのアプローチを概観するとともに, 協調フィルタリングやマルチタスク学習といった, 他の機械学習問題との関連を述べる.

2. ネットワーク構造データ

この章では, まず, 本解説で対象とするネットワーク構造データの定義について述べる. ネットワーク構造データとは, データとデータの間の関係を, グラフ構造で表現したものである. グラフのそれぞれのノードは, 1 つのデータを表し, リンクは, これが結ぶ 2 つのデータの間に何らかの関係があることを示している. リンクには向きがある場合と向きがない, あるいは何らかのラベル付けがなされている場合が考えられる.

例えば, 近年流行の SNS（ソーシャルネットワークサービス）を考えてみよう. SNS におけるネットワーク構造では, 各ノードがサービスの参加者, すなわち 1 人の人間を表し, リンクは, 2 人の参加者の間にお友達関係があることを表している (図 1).

表 1 に代表的なネットワーク構造データを纏めた.

ノードは必ずしも参加者だけではなく, コミュニティなどのグループをノードとして表わす場合もありうる. この場合, リンクはグループ間の関係や, 参加者がそのグ

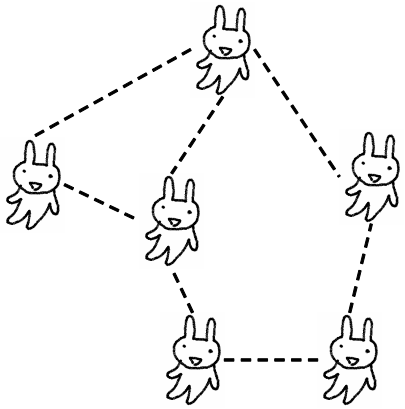

図 1 SNS（ソーシャルネットワークサービス）における社会ネットワークを表したグラフ構造．ノードは参加者，リンクは友人関係を表している．

ネットワーク	ノード	リンク
WWW	Web ページ	ハイパーリンク
社会ネットワーク	人間	人間関係
文献ネットワーク	論文	引用関係
生体ネットワーク	遺伝子 タンパク質	制御関係 相互作用

表 1 ネットワーク構造データの例

ループに属しているという関係を表すことになる*1．また，友人関係などの比較的静的な関係だけでなく，e メールの送信や，共同作業などの動的な環境をリンクとして表すこともありうる．この場合，グラフ構造も静的なものではなく，時間とともに構造が移り変わっていくと考えるべきだろう．

3. ネットワーク構造予測問題

3·1 リンクマイニング

ネットワーク構造をもったデータは，従来の機械学習やデータマイニングの分野ではその取り扱いの難しさ故，あまり扱われてこなかった対象であったが，近年のネットワーク構造データの増加による解析の必要性の高まりと，構造データを扱う手法の発展に伴い，にわかに注目を集めるようになった．データマイニングの分野では，ネットワーク構造データの解析は「リンクマイニング」と呼ばれ，現在も精力的に研究が進められている．この分野の第一人者である Getoor ら [Getoor 05] によれば，リンクマイニングの基本的なタスクには以下のようなものがある．

- ノードに関連するタスク
  - ノードのランキング
  - ノードの分類
  - ノードのクラスタリング
- 構造に関連するタスク
  - リンク予測（構造予測）
  - 構造パタン発見

どの問題の重要性も高く，様々な話題があるが，本稿では，ネットワーク構造の予測問題である，リンク予測問題に注目し，この問題への種々のアプローチを観ていくことにする．

3·2 リンク予測問題

本稿で扱うリンク予測問題を（緩やかに）定義しよう．リンク予測問題にはいくつかの設定がありうるが，大まかには「ネットワーク構造の既知の部分が与えられたとき，これを手がかりに未知の部分を予測する」と定義することができる．

例えば，実験的に相互作用があることが分かっているタンパク質のペアの間にリンクを張ることで構成されたネットワークの一部分を手がかりに，まだ知られていない相互作用を予測する問題を考えるとすると，部分的に分かっているネットワーク構造を手がかりにそれ以外の部分を予測するという「ネットワークの補完・外挿問題」と捉えることができる．この場合，新たに見込みがあると予測された相互作用の候補を，実際に実験的に確認してみることで，未知の相互作用を効率よく発見することが期待できる．

あるいは，別の例として，現在の人間関係を表した社会ネットワークが与えられたとき，今後，この人間関係がどのように変化していくかを予測する問題を考えると，予測をもとにして，SNS などにおける推薦機能を実現できるかもしれない．この問題は，「現在のネットワーク構造が与えられたとき，将来のネットワーク構造を予測する問題」と捉えることができる．場合によっては，これまでどのような過程を辿って構造が変化してきたかという，ネットワーク構造の遷移の履歴がデータとして利用できるような問題設定も考えられるだろう．

ネットワーク構造の既知の部分の与えられ方としても，2 つの場合が考えられる．リンクがあると分かっている場所と分かっていない場所が与えられる場合，つまり，教師あり学習の文脈で言えば，正例と負例が分かっている場合と，リンクがあると分かっている場所だけが与えられている場合，つまり，正例のみからの学習を行わなければならない場合の 2 種類の可能性がある．

4. リンク予測問題へのアプローチ

この章では，前章で紹介したネットワーク構造予測問題としてのリンク予測問題に対する種々のアプローチを紹介する．尚，本稿では，特に表記のない限り，「正例と負例の両方が与えられている，ネットワークの補完・外挿問題」を念頭において解説を進めるが，基本的な考え

*1 もっと一般的には，ネットワークをハイパーグラフとして表すことで，グループはハイパー辺として表現される．

方は, 他の設定にも比較的容易に拡張できる[*2].

4·1 教師付き学習としてのリンク予測

さて, ネットワーク構造の予測を行う際にも, (1) 各リンクの有無を別々に（独立に）予測するか, それとも (2) 構造全体を一気に（枝の有無が互いに依存していると考え）予測するかというアプローチの違いがある. ひとまずのところは, 前者 (1) の設定を考えることにしよう.

さて, 各リンクの有無は独立に予測してよいとするならば, リンク予測の問題は, 基本的に, ノード ペア の 2 値分類問題（2 つのノードの間にリンクが「ある」か「ない」か）か, ランキングの問題（2 つのノードの間にどのくらいリンクがありそうか）として考えることができる. つまり, 通常の機械学習アルゴリズムが, ノード, すなわち 1 つのデータのもつ特徴ベクトルに基づいて, そのデータがもつ性質についての予測を行っているのに対し, リンク予測では, ノードペア, すなわち 2 つのデータの特徴ベクトルに基づいて, 2 つのノード間の関係がもつ性質についての予測を行っていることになる. 従って, ノードペアに対して, 何らかの特徴ベクトルを定義できれば, リンク予測問題を, 通常の教師付き機械学習問題と同じような扱いかたによって解くことができるというわけである.

4·2 リンク予測に使うことのできる情報

リンク予測に用いることのできる情報は, 大きく分けて,「ノード関連の情報」と「構造関連の情報」に分けることができる.

- ノード関連の情報

  ノード情報とは, ノード自身が持っている情報であり, 例えば, SNS などにおいては参加者の個人情報, たとえば, 住所や年齢, 趣味などがあげられる. タンパク質の相互作用ネットワークにおいては, タンパク質の配列情報や局在部位情報, あるいは, 発現情報などがこれにあたる. これらは通常, 組み合わせてノード対の情報として用いられる.
- 構造関連の情報

  一方, 構造情報とは, リンク指標とも呼ばれ, 対象としているノードの周辺のリンク構造を捉えた情報であり, 古くは社会ネットワーク解析の分野から, 最近では, 情報検索の分野でも種々の指標が提案されている. 例えば, 「友達の友達が友達である確率が高い」という観察から, 2 つのリンクがつながっている場合, これらをショートカットするリンクが作られやすいというリンク指標が考えられる.

これらの情報は, その性質の違いから, 異なった取り扱いをする必要がある. 以下では, まず, これら 2 つの情報に基づくアプローチをそれぞれ紹介していくことにする.

*2 容易でない場合には, それが新しい研究テーマになったりするかもしれない.

4·3 ノード情報に基づくアプローチ

ネットワークの各ノード $v^{(i)}$ が, 特徴ベクトル $\mathbf{x}^{(i)}$ を持っているものとしよう. いまノードペア $v^{(i)}$ と $v^{(j)}$ に対し, 対応する特徴ベクトル $\mathbf{z}^{(i,j)}$ を, $\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ を用いてうまく定義したい. 簡単に思いつくものとしては, $\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ をつなげて大きな特徴ベクトルをつくるやり方が思い浮かぶが, これでは, $v^{(i)}$ と $v^{(j)}$ の情報を単純に足し合わせただけで, $v^{(i)}$ と $v^{(j)}$ の関係を表した特徴とは言い難い. そこで, ノードペアの特徴ベクトルを

$$\begin{aligned}\mathbf{z}^{(i,j)} &:= \mathbf{x}^{(i)} \otimes \mathbf{x}^{(j)} \\ &= \big(x_1^{(i)}x_1^{(j)}, x_1^{(i)}x_2^{(j)}, x_1^{(i)}x_3^{(j)}, \ldots, \\ &\quad x_2^{(i)}x_1^{(j)}, x_2^{(i)}x_2^{(j)}, x_2^{(i)}x_3^{(j)}, \ldots, \\ &\quad x_3^{(i)}x_1^{(j)}, x_3^{(i)}x_2^{(j)}, x_3^{(i)}x_3^{(j)}, \ldots \big)\end{aligned}$$

のように定義する [Oyama 04, Ben-Hur 05]. ここで $\otimes$ は, 直積を表すとする. つまり, ノードペアの特徴ベクトルを, もともとの特徴ベクトルの要素の組み合わせによって表現しようというのである. これなら, 片方のノードがある特徴を持っていて, もう一方のノードが別のある特徴を持っているという関係が表現できる. たとえば, タンパク質の相互作用なら, 片方のタンパク質がある特徴的な部分構造をもっていて, もう片方のタンパク質がそれに作用する別の部分構造を持っているといった具合である. 新しく作られた特徴ベクトル $\mathbf{z}^{(i,j)}$ の次元は, ノードの特徴ベクトルを D とすれば D^2 となる.

予測の際には, 例えば線形の予測器を用いるのならば, ノードペア $(v^{(i)}, v^{(j)})$ に対する予測は, パラメータを $\mathbf{w}$ として, 内積

$$h(v^{(i)}, v^{(j)}) = \langle \mathbf{w}, \mathbf{z}^{(i,j)} \rangle$$

によって予測を行う. ランキングであるなら, この値がリンクの存在しやすさとなり, 2 値分類であるなら符号の正負によってリンクの有無を予測することになる[*3]. 2 つのノードの関係は対称ではない, つまり, $\mathbf{z}^{(i,j)} \neq \mathbf{z}^{(j,i)}$ であることに注意する. つまり, ネットワーク構造が有向グラフである場合にはこのままでよいが, 無向グラフのとき, つまり, 関係に対称性が必要であるときには,

$$\bar{\mathbf{z}}^{(i,j)} := \mathbf{z}^{(i,j)} + \mathbf{z}^{(j,i)}$$

のように対称化してから使う必要がある.

さて, 前述したようにノードペアの特徴ベクトルの次元は $O(D^2)$ なので, テキストデータや, バイオインフォマティクスのデータなど特徴ベクトルの次元が高い場合にはあまり効率的ではない. そこで, 特徴ベクトルの次元によらない機械学習法であるカーネル法 [Shawe-Taylor 04] によるアプローチを考えよう. カーネル法では, カー

*3 モデルパラメータ $\mathbf{w}$ は, リンクの有無が分かっているノードペアから, 任意の教師付き学習アルゴリズムによって推定する.

ネル関数（2 つのデータの特徴ベクトルの内積で定義される）によって全てのデータアクセスを行うが, 我々の場合だと, 2 組のノードペア $(v^{(i)}, v^{(j)})$ と $(v^{(k)}, v^{(l)})$ に対する特徴ベクトルの内積がカーネル関数となる. つまり,

$$K((i,j),(k,l)) = \langle \mathbf{z}^{(i,j)}, \mathbf{z}^{(k,l)} \rangle$$

によってカーネル関数 K が定義される. すぐに確かめられるように, これは

$$\begin{aligned} \langle \mathbf{z}^{(i,j)}, \mathbf{z}^{(k,l)} \rangle &= \langle \mathbf{x}^{(i)}, \mathbf{x}^{(k)} \rangle \langle \mathbf{x}^{(j)}, \mathbf{z}^{(l)} \rangle \\ &= k(i,k)k(j,l) \end{aligned} \tag{1}$$

のようにノード同士のカーネル関数 k（ノードの特徴ベクトルの内積）の積として分解できるため, 実際の計算は, ノードの特徴空間の次元数のオーダーで行うことができる. カーネル法の場合には, ノードペア (i,j) に対する予測は, パラメータを α(ノードペア $(v^{(i)}, v^{(j)})$ に対する重み $\alpha^{(i)}$) として,

$$h(v^{(i)}, v^{(j)}) = \sum_{(k,l)} \alpha^{(k,l)} K((i,j),(k,l))$$

によって行う. この場合, ノード数を N とすると, パラメータ数が $O(N^2)$ になる. つまり, 特徴空間の次元とノード数のどちらが大きいかによって, カーネル法の効率のよさが決まるといえる.

また, カーネル法でないときと同様に, 無向グラフのときには

$$\bar{K}((i,j),(k,l)) = K((i,j),(k,l)) + K((j,i),(k,l))$$

とすればカーネル関数を対称化することができる.

このカーネルを用いて Oyama ら [Oyama 04] は文献データにおける同一著者の同定に, Ben-Hur ら [Ben-Hur 05] はタンパク質相互作用の予測に用い, 良好な性能を得ている.

さて, ノードペアの特徴ベクトルをそのまま使ったときにも, カーネル化した場合にも, 問題サイズが, ノード特徴の次元あるいはデータ数について 2 乗のオーダーになってしまう. これを解決するために, Vert ら [Vert 04] は, $\mathbf{z}^{(i,j)}$ の定義においてノードペア $(v^{(i)}, v^{(j)})$ の全ての特徴間の関係を直接考慮するのではなく, $\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ を一旦低次元の潜在変数ベクトル $\mathbf{f}^{(i)}$ と $\mathbf{f}^{(j)}$ に落とし, その空間での相関によってリンクの有無を予測する方法を提案した.

$\mathbf{x}^{(i)}$ を潜在変数による特徴空間においたときの, $\mathbf{f}^{(i)}$ の d 次元目の特徴 $f_d(\mathbf{x}^{(i)})$ は,

$$f_d(\mathbf{x}^{(i)}) = \langle \mathbf{v}_i, \mathbf{x}^{(i)} \rangle \tag{2}$$

あるいは, カーネル法の場合には,

$$f_d(\mathbf{x}^{(i)}) = \sum_j \beta_i^{(j)} \langle \mathbf{x}^{(j)}, \mathbf{x}^{(i)} \rangle \tag{3}$$

のように定義するとする. ここで $\mathbf{v}_i$ や $\beta^{(j)}$ はそれぞれ潜在変数への写像のためのパラメータである.

2 つのノード $(v^{(i)}, v^{(j)}$ の間にどのくらいリンクがなさそうか という値を, 新しい特徴空間における距離

$$\mathrm{dist}(v^{(i)}, v^{(j)}) = \sum_d (f_d(\mathbf{x}^{(i)}) - f_d(\mathbf{x}^{(j)}))^2$$

によって定義するとしよう. リンクの有無が既知であるノードペア $(v^{(i)}, v^{(j)})$ については, リンクが存在するならば $\mathrm{dist}(v^{(i)}, v^{(j)})$ が小さく, リンクが存在しないならば $\mathrm{dist}(v^{(i)}, v^{(j)}$ が大きくなるようにパラメータを学習すればよいことになる. これを, 各 f_d が直交するような制約の下で解くことで, 行列の固有値問題となり, 最適なパラメータを求めることができる. このとき, パラメータの数は, 潜在変数の数を H として DH, カーネル法の場合には NH となり, もともとのパラメータ数に比較して大きく減少していることがわかる.

4·4 構造情報に基づくアプローチ

ノード情報に基づくアプローチが各ノードのもつ情報を組み合わせて, ノードペアに対する特徴ベクトルを定義するのとは異なり, 構造情報に基づく特徴は 2 つのノードの周辺リンク構造に基づいて定義される. これらは, 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, 多くは社会ネットワーク分析の文脈で, ネットワーク構造の進化モデルを土台として提案されたものであるが, 近年では情報検索の分野でも提案されている.

以下, これらの指標をリンク指標と呼ぶことにし, いくつかの具体例を紹介しよう. なお, $\Gamma(v^{(i)})$ をノード $v^{(i)}$ の隣接ノードの集合 (ノード $v^{(i)}$ に連結したノードの集合) とする.

- 共通隣接ノード (common neighbors) 指標 [Newman 01]

  $$\mathrm{common}(v^{(i)}, v^{(j)}) := |\Gamma(v^{(i)})| \cap |\Gamma(v^{(j)})|$$

  共通隣接ノード指標はノード $v^{(i)}$ とノード $v^{(j)}$ 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. 直感的には, これは「共通の友人が多い 2 人が, 互いに友人である可能性は高いだろう」という仮説を表現したものである. 研究者の共同研究のネットワークなどが, この規則に従う傾向が強いようである.
- Jaccard 係数 [Baeza-Yates 99, Liben-Nowell 04]

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

  Jaccard 係数は, 情報検索の分野では類似度としても用いられる指標である. 定義を見てもわかるように, これは正規化された共通隣接ノード指標になっており, 共通の隣接ノードが, 2 つのノードの隣接ノード

集合に占める割合を示している．直感的には「友達の大半が重なるのなら, 2 人は友達である可能性が高いだろう」といった具合である．

- Adamic/Adar [Adamic 03]

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

Adamic/Adar 指標は, 重みつきの共通隣接ノード指標であり, 隣接ノードごとに異なった重みが割り当てられる．重みの大きさは, 隣接ノードがもつ隣接ノードの数に応じて決められ, 少数の隣接ノードをもつノードを共通隣接ノードとして持つと, 大きな重みが割り当てられる．つまり「人付き合いの悪い人を共通に友達にもつ 2 人は, 友達である可能性が高いだろう」という感じである．

- Katz_β 指標 [Katz 53]

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

Katz_β 指標は共通隣接ノード指標の一般化であり, 隣だけでなく, より遠くの関係をも考慮する．$\mathrm{paths}^{(l)}_{v^{(i)}, v^{(j)}}$ はノード $v^{(i)}$ からノード $v^{(j)}$ への長さ l のパスの集合を表すとする．隣接ノードが 1 ステップで到達できるノードであるとすると, 2 ステップ以上で到達できるノードは隣接ノードの自然な一般化になっており, 従って, 2 つのノードを結ぶパスの集合の数は, 共通隣接ノード数を一般化したものと考えられることができる．Katz_β 指標は, 2 つのノードを結ぶパスの集合の数の重みつき和（β は 0 以上 1 未満の減衰係数）をとったものであり, 重みはパスの長さに応じて指数的に減衰する．ちなみに, この指標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] と同じ構造をもっており, 両者は非常に類似している．

- 優先的選択指標 [Newman 01, Barabási 02]

$$\mathrm{preferential}(v^{(i)}, v^{(j)}) := |\Gamma(v^{(i)})| \cdot |\Gamma(v^{(j)})|$$

優先的選択 (preferential attachment) 指標は Barabási によって提案された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り, 上記の指標とは若干異なっている．「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」というモデルに基づくこの指標では, リンクの得られやすさは, 2 つのノードのもつ隣接ノード数の掛け算で定義され, 必ずしも共通の隣接ノードをもっている必要はない．

ここで紹介した各指標は, 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いるので, ノードペアのランキングを与えることになる．従って, 最も簡単なリンク予測法としては, 全てのノードペアをリンク指標の大きい順に並べて, 適当な閾値以上のペアの間にリンクがあると予測する方法である [Liben-Nowell 04].

どの指標がリンク予測に適しているかという問いには一般的な答えは無く, 予測性能の良し悪しは, 対象とするデータが, 各指標が想定しているネットワークの生成モデルに, どれだけマッチしているかに大きく依存する．

従って, これらの指標を特徴ベクトルの要素として学習アルゴリズムに渡すことによって, 予測精度に応じた重み付けがなされることになる．（通常は, 次節で述べるように, ノードの特徴と組み合わせて用いられる．）

なお, ここではリンク指標は通常, 無向グラフについて定義されているが, 有向グラフの場合にはリンクの向きごとに別々に扱うなどの工夫が必要になるだろう．

4·5 ノード情報と構造情報を組み合わせる

これまで, ノード情報と構造情報のどちらかに基づいた予測手法を紹介したが, 実際にこれらの手法を適用する場面では, 両方の情報が与えられている場合が多い．例えば, タンパク質相互作用ネットワークでは, 個々のタンパク質の情報がノード情報として, 相互作用の情報が構造情報として与えられる．

最も単純に 2 種類の情報を組み合わせる方法としては, それぞれの特徴ベクトルを連結し, 教師あり学習アルゴリズムに渡すというやり方が考えられる [Hasan 05, O'Madadhain 05]．この場合, ノード情報に基づく特徴ベクトル $\mathbf{z}^{(i,j)}$ に, 構造情報に基づくリンク指標が新しい次元として付加される．

カーネル法の場合には, ノード情報に基づくカーネル関数と, 構造情報に基づくカーネル関数, 例えば, Katz_β 指標に類似している拡散カーネルを組み合わせることになる．組み合わせは, 例えば 2 つの情報に基づくカーネル関数の線形結合 [Basilico 04] となり,

$$K((i,j),(k,l)) = \mu_V K_V((i,j),(k,l)) + \mu_S K_S((i,j),(k,l))$$

のように, ノード情報に基づくノードペア同士のカーネル K_V と, 構造情報にもとづくノードペア同士のカーネル K_S（ともに (1) によって定義する）の重みつき線形和として表すこともできる[*4]．この場合, 最適な混合割合（μ_V と μ_S）を決定するのは半正定値計画法などの比較的難しい最適化問題を解くことが必要になる [Bach 04, Lanckriet 04].

さて, 以上の方法は, 訓練データ（観測された部分）とテストデータ（未知の部分）の両方において, ノード情報と構造情報の両方の種類のデータが得られることを想定している．しかしながら, 訓練データには両方とも存在しているが, テストデータには片方しか存在していない場合がある．例えば, タンパク質の相互作用ネットワー

[*4] 2 つ以上のカーネル関数の線型和でもよい.

クにおいては, 訓練データに関しては, 両方の情報が得られるが, テストデータに関しては, ほぼノード情報のみしか与えられてないという状況が存在する. この場合には, 予測時には構造情報を用いることができず, ノード情報のみからリンク予測を行う必要がでてくる.

この問題に対する解決策として, Yamanishi らは, 訓練時には両方の情報を用い, 予測時にはノード情報のみを用いる手法を提案した [Yamanishi 05]. 基本的には,「ノード情報に基づくアプローチ」の節で紹介した, 低次元空間への写像を用いた予測を行うが, 単に, リンクのあるノードペアについては 2 つのノードの潜在変数の距離が小さく, 無いペアについては距離が大きくなるように学習を行うのではなく, リンク構造から導かれたリンク指標に, 潜在変数の相関の値が近づくように学習を行う. 予測時に構造情報を用いることができないので, 構造情報に基づく予測を, ノード情報によって再現できるように学習を行っている点がポイントである. 彼らは拡散カーネル [Kondor 02] をリンク指標として用い, 最適化問題を（カーネル）正準相関分析として定義することで, タンパク質相互作用ネットワークの予測を行っている.

なお, Kato ら [Kato 05] は, 様々なノード情報に基づく複数のカーネル関数を統合するために, これらの線型結合を, 拡散カーネルの値を近似するように結合重みを学習することによって, ネットワークの未知の部分を予測する方法を提案している.

4·6 ネットワーク構造全体の確率モデル

この章のはじめで述べたように, ここまでのところ, 各リンクの有無は別々に予測してよいとしてきた. しかし, この仮定はどちらかというと計算上の都合であって, 本来ならば全てのリンクの同時分布を考えるべきところを, 各リンクについての周辺分布の積で考える平均場近似のような近似を行っていることになる. この仮定によって, ネットワーク構造全体の予測問題を, ノードペアについての予測問題に単純化して扱えたのであった. しかし, ちゃんと考えてみると, 構造情報は, ノードペアの周りのリンク構造であったので, あるリンクの有無についての予測は, その周りのリンク予測に影響を与えると同時に影響も受けているはずである. また, ノード情報にしても, 必ずしもそのノードが両端のどちらかに含まれるリンクの有無にしか影響を与えないわけではないであろう. つまり, ネットワーク構造には, ノード情報も含む, 何らかの局所的な構造パターン[*5]が存在して, ネットワーク構造はこれらの組み合わせによって形作られていると考えることができる. このような場合には, ネットワーク構造全体の整合性を考えて予測する, 構造全体のモデルを考える必要がある.

Taskar ら [Taskar 03] は, ネットワークの既知の部分を G^L, 未知の部分を G^U として, ネットワーク構造全体に対する特徴ベクトル $\boldsymbol{\phi}(G^U, G^L)$ を考え,

$$G^U = \underset{G^U}{\operatorname{argmax}} \langle \mathbf{w}, \boldsymbol{\phi}(G^U, G^L) \rangle$$

によって構造を予測するというモデルを考えた. ここで, $\boldsymbol{\phi}$ の各要素は, 前述の「局所的な構造パターン」がネットワーク構造内に何回出現するかを表すものとする. つまり, ネットワークの既知の部分 G^L とパラメータ $\mathbf{w}$ が与えられたときに, 評価値 $\langle \mathbf{w}, \boldsymbol{\phi}(G^U, G^L) \rangle$ を最も大きくするような G^U を決定することになる. G^U が変わると, $\boldsymbol{\phi}(G^U, G^L)$ も変わることに注意する.

パラメータ $\mathbf{w}$ の推定方法として, Taskar らは次の指数分布族型の確率モデル

$$h(G^U | G^L) = \frac{\exp(\langle \mathbf{w}, \boldsymbol{\phi}(G^U, G^L) \rangle)}{\sum_{G^U} \exp(\langle \mathbf{w}, \boldsymbol{\phi}(G^U, G^L) \rangle}$$

を考え, 最尤推定によるパラメータ推定法を提案している [*6].

Taskar らのモデルは, 非常に汎用的な枠組みである, 確率的な関係学習モデルであり, 静的なネットワークのモデルとしては, ほぼ決定的なモデルといってもよいモデルであるが, 学習と予測ともに, 指数的に多い候補の中から最良のものを選び出してきたり, 全ての候補についての和をとったりする必要があるため, 実際の適用においては, サンプリングなどによって近似的に解を求めるなどの計算上の妥協が必要になる. このモデルが, リンク予測の独立性を仮定する方法と比較して, 精度や速度実などの面においてどれほどよいかどうかについては, 今後の検証が待たれるところであろう.

ところで, 近年, 世の中の多くのネットワーク構造は「スケールフリー性」, すなわち, ノードの次数の分布がベキ分布に従っているという性質をもつことが示されている. リンク指標の節で紹介した優先的選択指標もスケールフリーネットワークの生成モデルを基礎としているが, もっと直接的に, 予測にスケールフリー性を取り入れたのが Gomez ら [Gomez 01] である.

Gomez らは予測の際に, 予測されるネットワーク構造の候補がもつ確率スコアに, 各ノードがもつリンクの次数に応じたベキ分布 (図 2) の確率を掛け合わせることで, よりスケールフリー性の高いネットワークが予測されるようにバイアスをかけるというアプローチを提案している.

もちろん機械学習の立場からすると, このようなバイアスは, 学習時に正則化あるいは事前分布の形で入れるのが望ましいものの, 彼らのモデルはスケールフリーネットワークとネットワーク構造予測の融合を計った先駆的な試みといえる.

---

*5 ネットワークモチーフなどと呼ばれたりもする.

*6 実は, 社会ネットワーク解析の分野でも, p^* モデル [Anderson 99] と呼ばれる同様のモデルがすでに提案されていたりする.

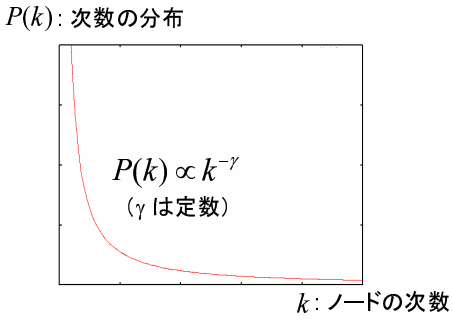

図 2 スケールフリーネットワークでは, ノードの次数がベキ分布に従う.

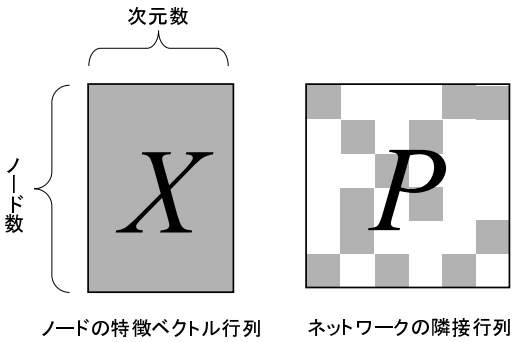

図 3 抽象化されたリンク予測の問題設定. 灰色部分が訓練データとして与えられる部分.

5. 他の問題との関係

ここまでは主に, リンク予測を目的として開発された手法を紹介してきたが, 実は, リンク予測問題は, 他の種類の学習問題と非常に類似した構造を持っている. 従って, データの与えられ方の仮定などの問題固有の性質の違いはあれど, 抽象化されたレベルにおいては, 他の問題に対して設計された手法が, リンク予測の問題に自然に適用できるといえる. また同様に, リンク予測の方法が, 他の問題に対して適用できることもありうる.

リンク予測の問題は, 基本的には, 図 3 のように, 付加情報としてノードの特徴ベクトル行列 X が与えられたときの, グラフの隣接行列 P の穴埋め問題として考えることができる. (場合によっては X は与えられない場合もある. また, X の代わりにカーネル行列が与えられる場合もある.)

この問題は, P が正方行列でなく, 各行が商品などのアイテム, 各列がユーザーに対応していると考えると, 協調フィルタリング[*7]の問題であると考えることもできる. 協調フィルタリングは, 各ユーザーの各アイテムへの評価をもとに, 未知の評価を予測する問題であり, 協調フィルタリングにおいてもやはり, P のみを用いる場合と, アイテムの特徴ベクトル行列 X, それにユーザーの特徴ベクトル Y が与えられる場合が考えられる. 従って, 協調フィルタリングの手法を比較的簡単にリンク予測の方法に適用することも可能であるし, その逆もありうる. 例えば, Huang ら [Huang 05] は, リンク指標に基づく予測を, 協調フィルタリングに適用するという試みを行っている. また, Kashima ら [Kashima 06, 鹿島 06] は, 重みつき多数決に基づく協調フィルタリングの手法 [Nakamura 98] を, リンク予測の手法として拡張するとともに, ネットワークの構造進化モデル的解釈を与えている.

また, 複数の関連ある教師付きタスクを, 別々に解くのではなく, タスク間の類似性に基づきお互いのデータを利用しながら, 一度に解くというマルチタスク学習問題 [Caruana 97] も同様の構造をもっている. この場合, P の行がタスクに, 列が事例に相当することになる. マルチタスク学習では通常, リンク予測や協調フィルタリングにおける X や Y に対応するものはないが, タスクが特徴を持つような場合も当然考えられるだろう.

いずれの問題に対しても, 本解説で紹介したような潜在的な変数の存在を仮定したようなモデルが提案されており, それらの関係も興味深い [Yu 05].

[*7] 協調フィルタリングについては http://www.ai-gakkai.or.jp/jsai/whatsai/AItopics2.html に, 種々の手法については, 麻生らによるサーベイ [麻生 06] が詳しい.

6. おわりに

本稿では, ネットワーク構造の予測問題へのいくつかのアプローチを概観した. 今後の発展としては, より高度なモデル化と手法の適用や, ネットワーク構造の時間変化モデルなどがあるだろう.

近年ではこの分野でもベイズ的なアプローチが盛んである. 例えば, [Chu 07, Yu 07] は, Vert ら [Vert 04] のアプローチの延長線上にある, 潜在変数 f がガウス過程によって生成されるようなモデルを提案している.

また, ネットワーク構造の時間変化の履歴が与えられ, そこから構造変化のダイナミクスを学習するという問題設定も興味深い. 例えば, 4･6 節で紹介したようなネットワーク構造全体のモデル [Taskar 03, Anderson 99] を, 時間方向に拡張するようなモデル [Hanneke 06] も提案されている.

◇ 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)

[Anderson 99] Anderson, C. J., Wasserman, S., and Crouch, B.: A p^* primer: logit models for social networks, *Social Networks*, Vol. 21, pp. 37–66 (1999)

[Bach 04] Bach, F. R., Lanckriet, G. R. G., and Jordan, M. I.: Multiple Kernel Learning, Conic Duality, and the SMO Algorithm, in *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)* (2004)

[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)
[Basilico 04] Basilico, J. and Hofmann, T.: Unifying collaborative and content-based filtering, in *In Proceedings of the Twenty-first International Conference on Machine Learning (ICML)* (2004)
[Ben-Hur 05] Ben-Hur, A. and Noble, W. S.: Kernel methods for predicting protein-protein interactions, *Bioinformatics*, Vol. 21, No. Suppl. 1, pp. i38–i46 (2005)
[Caruana 97] Caruana, R.: Multitask Learning, *Machine Learning*, Vol. 28, No. 1, pp. 41–75 (1997)
[Chu 07] Chu, W., Sindhwani, V., Ghahramani, Z., and Keerthi, S.: Relational Learning with Gaussian Processes, in Schölkopf, B., Platt, J., and Hoffman, T. eds., *Advances in Neural Information Processing Systems 19*, MIT Press, Cambridge, MA (2007)
[Getoor 05] Getoor, L. and Diehl, C. P.: Link mining: a survey, *SIGKDD Explorations*, Vol. 7, No. 2, pp. 3–12 (2005)
[Gomez 01] Gomez, S. M., Lo, S.-H., and Rzhetsky, A.: Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks, *Genetics*, Vol. 159, pp. 1291–1298 (2001)
[Hanneke 06] Hanneke, S. and Xing, E.: Discrete Temporal Models of Social Networks, in *the Workshop on Statistical Network Analysis, held at the 23rd International Conference on Machine Learning* (2006)
[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)
[Huang 05] Huang, Z., Li, X., and Chen, H.: Link Prediction Approach to Collaborative Filtering, in *Proceedings of the Fifth ACM/IEEE-CS joint conference on Digital libraries (JCDL)* (2005)
[Kashima 06] Kashima, H. and Abe, N.: A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction, in *Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM)* (2006)
[Kato 05] Kato, T., Tsuda, K., and Asai, K.: Selective integration of multiple biological data for supervised network inference, *Bioinformatics*, Vol. 21, No. 5, pp. 2488–2945 (2005)
[Katz 53] Katz, L.: A new status index derived from sociometric analysis, *Psycometrika*, Vol. 18, No. 1, pp. 39–43 (1953)
[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)
[Lanckriet 04] Lanckriet, G. R. G., Deng, M., Cristianini, N., Jordan, M. I., and Noble, W. S.: Kernel-Based Data Fusion and Its Application to Protein Function Prediction in Yeast, in *Proceedings of the Pacific Symposium on Biocomputing (PSB)* (2004)
[Liben-Nowell 04] 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)
[Nakamura 98] Nakamura, A. and Abe, N.: Collaborative filtering using weighted majority prediction algorithms, in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, pp. 395–403 (1998)
[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)
[Oyama 04] Oyama, S. and Manning, C. D.: Using Feature Conjunctions across Examples for Learning Pairwise Classifiers, in *Proceedings of the Fifteenth European Conference on Machine Learning (ECML)*, pp. 322–333 (2004)
[Shawe-Taylor 04] Shawe-Taylor, J. and Cristianini, N.: *Kernel Methods for Pattern Analysis*, Cambridge University Press (2004)
[Taskar 03] Taskar, B., Wong, M., Abbeel, P., and Koller, D.: Link prediction in relational data, in *Neural Information Processing System* (2003)
[Vert 04] Vert, J.-P. and Yamanishi, Y.: Supervised Graph Inference, in *Proceedings of the Neural Information Processing 2004 (NIPS)* (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)
[Yu 05] Yu, K. and Tresp, V.: Learning to Learn and Collaborative Filtering, NIPS 2005 workshop " Inductive Transfer: 10 Years Later " (2005)
[Yu 07] Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z.: Stochastic Relational Models for Discriminative Link Prediction, in Schölkopf, B., Platt, J., and Hoffman, T. eds., *Advances in Neural Information Processing Systems 19*, MIT Press, Cambridge, MA (2007)
[鹿島 06] 鹿島 久嗣, 安倍 直樹：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, 人工知能学会論文誌, Vol. 22, No. 2 (2006)
[麻生 06] 麻生 英樹, 小野 智弘, 本村 陽一, 黒川 茂莉, 櫻井 彰人：協調フィルタリングと属性ベースフィルタリングの統合について, 電子情報通信学会研究報告 ニューロコンピューティング研究会 (NC), NC2006-54 (2006)

謝　　辞

本稿を執筆するにあたり, 貴重な情報やご意見を頂きました阿久津達也先生, 神嶌 敏弘先生, 村田 剛志先生, 山西芳裕先生, 上田展久先生に感謝いたします.

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, 現在に至る. 2007 年に京都大学情報学研究科知能情報学専攻にて博士後期課程修了. 博士 (情報学). 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.