無限潜在特徴関係モデルのマージン最大化推定による離散関係属性付きネットワークの分析

Analyzing Networks with Discrete Attributes using Max-Margin Nonparametric Latent Feature Relational Models

西出飛翔 † 江口浩二 †
Tsubasa NISHIDE Koji EGUCHI

1. はじめに

社会的ネットワークや生物学的ネットワークなどの関係データを利用する頻度や範囲が拡大すると共に，それらのデータを統計分析に充てることで有用な知見を得ることが課題となっている．これらのデータは一般的にエンティティをノードで，エンティティ間の関係をリンクで表すグラフ構造として表すことができる．リンク予測は社会的ネットワークや生物学的ネットワークなどの関係データの分析において基本的な問題の一つとして挙げられ，ノード間における既観測のリンクから未観測のリンクを予測する問題を指す [1]．既観測のリンクだけでなく，各エンティティの持つ属性などの情報も使用してリンク予測を行うこともある [2] [3]．

確率モデルに基づいた様々な手法によるリンク予測に関する研究は発展し続けている．そのようなモデルの中でも本稿で着目するものは，リンク構造の確率分布を定義するために各ノードが持つと仮定する潜在特徴行列と，シグモイド関数などのリンク尤度関数を利用するモデルである [3] [4]．

潜在特徴行列の未知の次元数を決定するためには，ほとんどの手法では交差検定などによるモデル選択が必要である．これは多くの異なる訓練データによる結果を比較する必要がある．そこで Miller らはインド料理過程 (Indian Buffet Process: IBP) [5] によるノンパラメトリックベイズ法を用いて，データから未知の潜在特徴の次元数を推定することを提案した [3]．これが無限潜在特徴関係モデル (Latent Feature Relational Model: LFRM) である．さらに LFRM を発展させて最大エントロピー識別 (Maximum Entropy Discrimination: MED) [6] [7] に基づいてヒンジ損失などの目的関数を直接最小化することでリンク予測を行うモデルが Zhu によって提案された [8]．これがマージン最大化潜在特徴関係モデル (Max-Margin Latent Feature Relational Model: MedLFRM) である．このモデルはノンパラメトリックベイズ法とマージン最大化法という二つ手法を統合したモデルである．ここでは，ベイズ推定の計算は目的関数を最小化することと等価になるため，教師データを用いることが可能となる．また，ソフトマージンにより誤分類を許容して柔軟なモデルを実現できる．ソフトマージン最大化に関する部分問題は，すでに存在する性能の高い解法によって解くことができる．

MedLFRM によって，LFRM よりもリンク予測の効率や精度が向上したことは示されている [8]．しかしながら，そこでは連続値として表現された関係属性のみ仮定されており，ソーシャルメディアにおいてしばしばみられる多次元離散関係属性については検討されていない．本稿ではこの点に着目した評価を行う．

本稿の実験は，LFRM や MedLFRM では連続値表現が仮定された関係属性を多次元離散表現であるものと仮定する．これは実問題のソーシャルメディアなどのデータセットにおいて関係属性は二値で表現されることが多いことを考慮したためである．これを踏まえて Facebook(注1) のデータセットを用いてリンク予測問題に対して実験を行ったところ，今後の研究に関して有用な知見を得た．

この論文の構成は以下の通りである．第 2 章では，LFRM, MED，MedLFRM といった既存手法を紹介する．第 3 章では，MedLFRM による離散関係属性付きネットワーク解析について述べ，その条件の下で MedLFRM のパラメータの推定方法について述べる．離散関係属性を考慮したリンク予測実験の結果を第 4 章で示し，最後に第 5 章で結論を述べる．

2. 関連研究

この章では関連研究して既存手法について述べる．初めにネットワークデータ (関係データ) のための潜在変数モデルである無限潜在特徴関係モデル (LFRM)，次に目的関数を最小化することで事後分布を学習する手法である最大エントロピー識別 (MED)，そしてそれらを組み合わせたマージン最大化潜在特徴関係モデル (MedLFRM) について示す．

2.1 無限潜在特徴関係モデル

潜在特徴関係モデル (LFRM) は Miller らによって提案されたものである [3]．このモデルは，各ノードが二値の潜在特徴の集合を持つと仮定し，その潜在特徴空間の次元をデータから推定すると同時にネットワークデータのリンクが生成される尤度を推定するモデルである．ノード間のリンクは各ノードの持つ潜在特徴と，ノード間の関係属性，そしてそれらの重みから生成される．

† 神戸大学, Kobe University

(注1)：http://www.facebook.com

ネットワーク内のノード数を N とし，$N \times N$ の二値隣接行列行列を Y とする．つまり，ノード i とノード j との間にリンクがある場合は $Y_{ij} = +1$ とし，そうでない場合は $Y_{ij} = -1$ とする．Y は完全に観測されてはおらず，既観測のリンクから未観測のリンクの値を予測できるモデルを学習することが目的である．また，ノード i とノード j との間のリンクに作用する関係属性 $X_{ij} \in \mathbb{R}^D$ が観測される場合もある．これは各成分が [0,1] の実数値ベクトルであり，ノード i とノード j のそれぞれに与えられた属性から求めることができる．たとえば，ソーシャルネットワークにおいて友人関係かどうかをリンクで表す問題を考えるとすると，X_{ij} は各ノード間の地理的特性を表すことができる．つまり，ユーザーが近い場所に住んでいる場合と，そうでない場合といった特性を X_{ij} の値によって表現することができる．

各ノードの持つ潜在特徴の数を K とすると，LFRM ではノードは二値潜在特徴ベクトル $\mu_i \in \mathbb{R}^K$ の集合で表される．ここで Z を $N \times K$ の二値潜在特徴行列とすると，$Z = [\mu_1^\mathsf{T}; ...; \mu_N^\mathsf{T}]$ である．Z_i はノード i の特徴ベクトルを表し，ノード i が特徴 k を持つ場合 $Z_{ik} = 1$，そうでない場合は $Z_{ik} = 0$ である．W を $K \times K$ の実数値重み行列とし，$W_{kk'}$ は，ノード i が特徴 k を持ちノード j が特徴 k' を持つならば，ノード i からノード j へのリンクに影響を与える重みであるとする．以上からリンク尤度は一般に以下のように定義される．

$$p(Y_{ij} = 1 | X_{ij}, \mu_i, \mu_j) = \Phi(\mu + \boldsymbol{\beta}^\mathsf{T} X_{ij} + \mu_i^\mathsf{T} W \mu_j) \qquad (1)$$

ここで Φ はシグモイド関数である．また，μ は尤度に影響を与える大域的バイアス値であり，$\boldsymbol{\beta}$ は関係属性の実数値重みベクトルである．

適した事前分布を得るために，LFRM ではインド料理過程 (IBP) [5] を Z の事前分布として使用する．これは Z を推定すると同時に，特徴数 K も推定するためである.W に関しては各成分において独立して事前分布 $N(0, \sigma_w^2)$ をとる．

IBP は非有界な二値行列の事前分布として用いられる．ここから得られる行列は，潜在特徴をいくつ持っていたとしても必ず各成分は正の値を取る．行列の成分 (各ノードの潜在特徴) のサンプリングは以下のように行われる．1 番目のノードに対応する行のうち，Poisson(α) だけの数の成分を 1 とする．ここで α はハイパーパラメータである．次に i 番目のノードに対応する行のうち，すでに他のノードに対応する行で 1 となっている成分は，その 1 となっている数に比例した確率で 1 となる．また，Poisson(α/i) だけの数の成分を新しく 1 にする．これを有限個のノード数だけ繰り返すことで潜在特徴行列を得る．この過程を Fig. 1 に示す．また，この過程は交換可能なので選択される行の順番は影響しない．

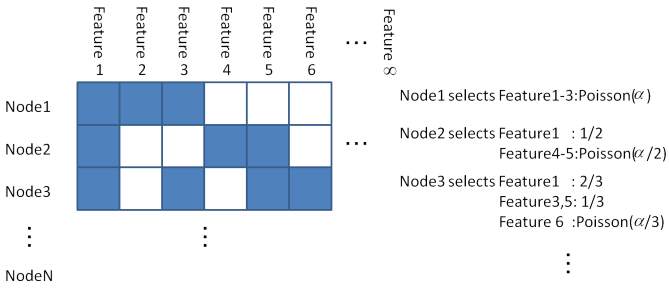

Fig. 1　Behavior of IBP.

2.2　最大エントロピー識別

最大エントロピー識別 (MED) [6] [7] は定義された目的関数である正則化項付き相対エントロピー最小化問題を解くことで事後分布を学習する手法である．

応答変数 Y が $\{+1, -1\}$ をとる二値分類問題を考えるとする．X を入力属性ベクトル，η をパラメータ，$F(X_n; \eta) = \eta^\mathsf{T} X_n$ を識別関数と定義する．また，ℓ を正の損失パラメータとして，ヒンジ損失関数を $h_\ell(x) = \max(0, \ell - x)$ と定義する．η の事前分布を $p_0(\eta)$，η の事後分布を $p(\eta)$ とすると，目的関数である正則化項付き相対エントロピー最小化問題は一般的に以下のように置くことができる．

$$\min_{p(\eta)} \ \mathrm{KL}(p(\eta)|p_0(\eta)) + C\mathcal{R}(p(\eta)) \qquad (2)$$

ここで C は正の定数，$\mathrm{KL}(p|q)$ は KL ダイバージェンス，すなわち相対エントロピーであり，$\mathcal{R}(p(\eta)) = \sum_n h_\ell(Y_n \mathbb{E}_{p(\eta)}[F(X_n; \eta)])$ はヒンジ損失である．ヒンジ損失は識別関数の予測結果と観測値との誤差の和に基づく．

応答変数 Y の予測値は以下のようになる．

$$\hat{Y} = \mathrm{sign}\mathbb{E}_{p(\eta)}[F(X; \eta)] \qquad (3)$$

2.3　マージン最大化潜在特徴関係モデル

マージン最大化潜在特徴関係モデル (MedLFRM) は Zhu によって提案された [8]．これは MED の識別関数 F を LFRM によるリンク尤度として定義することで，効率的にリンク尤度を推定することができるモデルである．

LFRM モデルと同様に二値潜在特徴行列 Z_i と Z_j，実数値重み行列 W，実数値関係属性 X_{ij} が与えられれば，識別関数は以下のようにして定義できる．

$$\begin{aligned} f(Z_i, Z_j; X_{ij}, W, \eta) &= Z_i W Z_j^\mathsf{T} + \eta^\mathsf{T} X_{ij} \\ &= \mathrm{Tr}(W Z_j^\mathsf{T} Z_i) + \eta^\mathsf{T} X_{ij} \end{aligned} \qquad (4)$$

η は関係属性の実数値重みベクトルである．ここで，$\Theta = \{W, \eta\}$ を全てのパラメータとする．Θ と事前分布 $p_0(\Theta)$ は確率変数とする．予測を行うための効果的な識別関数を $p(Z, \Theta)$ に関する期待値として定義すると以下のようになる．

$$f(X_{ij}) = \mathbb{E}_{p(Z,\Theta)}[f(Z_i, Z_j; X_{ij}, \Theta)] \qquad (5)$$

したがって，予測値は $\hat{Y_{ij}} = \mathrm{sign} f(X_{ij})$ である．$\mathcal{E}$ を観測されたリンクの組の集合とし，2.2 節と同様に $h_\ell = \max(0, \ell - x)$

と定義すると，ヒンジ損失は以下のようになる．

$$\mathcal{R}_\ell(p(Z,\Theta)) = \sum_{(i,j)\in\mathcal{E}} h_\ell(Y_{ij} f(X_{ij})) \tag{6}$$

$p_0(Z)$ を潜在特徴行列の事前分布とする．以上から，MedLFRM を以下の問題を解くことと定義することができる．

$$\min_{p(Z,\Theta)\in\mathcal{P}} \mathrm{KL}(p(Z,\Theta)|p_0(Z,\Theta)) + \mathcal{C}\mathcal{R}_\ell(p(Z,\Theta)) \tag{7}$$

一般的に，適宜補助的な変数を導入し，変数の依存関係を条件付き独立に変換することで推定を簡単化することがしばしば行われ，変分近似と呼ばれる．これを行うために Teh らによって提案された IBP のための棒折り過程 (Stick Breaking Prior: SBP) [9] を用いる．$\pi_k \in (0,1)$ を Z の列 k と対応付けられたパラメータとする．パラメータ $\boldsymbol{\pi} = \{\pi_k\}$ は棒折り過程によって生成され，$\pi_1 = \nu_1$，そして $\pi_k = \nu_k \pi_{k-1} = \prod_{i=1}^{k} \nu_i$ とする．ここで ν_i は α をハイパーパラメータとするベータ分布 $\mathrm{Beta}(\alpha, 1)$ からサンプリングされると仮定する．π_k が与えられると列 k の各 Z_{nk} は，π_k をハイパーパラメータとするベルヌーイ分布 $\mathrm{Bernoulli}(\pi_k)$ から独立してサンプリングされるとする．

MedLFRM において仮定するリンクの生成過程を以下に示す．

（1）潜在特徴行列の 1 行目である Z_1 に対して，
- （a）ポアソン分布 $\mathrm{Poisson}(\alpha)$ から成分数 M を選択
- （b）ベータ分布 $\mathrm{Beta}(\alpha, 1)$ から，各成分 $i \in \{1, ..., M\}$ に対して，パラメータ ν_i を選択
- （c）棒折り過程 $\mathrm{SBP}(\boldsymbol{\nu})$ から，各成分 $k \in \{1, ..., M\}$ に対して，パラメータ π_k を生成
- （d）ベルヌーイ分布 $\mathrm{Bernoulli}(\pi_k)$ から，各成分 $Z_{1k} \in \{Z_{11}, ..., Z_{1M}\}$ を選択

（2）潜在特徴行列の n 行目である Z_n に対して，
- （a）これまで未選択の成分に対してはポアソン分布 $\mathrm{Poisson}(\alpha/n)$ から，すでに選択されたことのある成分に対してはその確率から成分数 M を選択
- （b）ベータ分布 $\mathrm{Beta}(\alpha, 1)$ から，各成分 $i \in \{1, ..., M\}$ に対して，パラメータ ν_i を選択
- （c）棒折り過程 $\mathrm{SBP}(\boldsymbol{\nu})$ から，各成分 $k \in \{1, ..., M\}$ に対して，パラメータ π_k を生成
- （d）ベルヌーイ分布 $\mathrm{Bernoulli}(\pi_k)$ から，各成分 $Z_{nk} \in \{Z_{n1}, ..., Z_{nM}\}$ を選択

（3）正規分布 $N(0, \sigma_w^2)$ から，重み行列 W に対して，各成分を選択

（4）リンク評価関数 $\Phi(Z_i W Z_j^{\mathsf{T}} + \eta^{\mathsf{T}} X_{ij})$ によって，各ノード対 $(i,j) \in \mathcal{E}$ に対して，応答変数 Y_{ij} を生成

以下のように補助変数 $\boldsymbol{\nu}$ を導入したことによって拡張された MedLFRM の問題を得ることができる．

$$\min_{p(\boldsymbol{\nu},Z,\Theta)} \mathrm{KL}(p(\boldsymbol{\nu},Z,\Theta)|p_0(\boldsymbol{\nu},Z,\Theta)) + \mathcal{C}\mathcal{R}_\ell(p(Z,\Theta)) \tag{8}$$

ここで $p_0(\boldsymbol{\nu}, Z, \Theta) = p_0(\boldsymbol{\nu}) p(Z|\boldsymbol{\nu}) p_0(\Theta)$

3. 離散関係属性を伴う MedLFRM

前章ではネットワークデータのための潜在変数モデルである MedLFRM によるリンク予測問題について示してきた．この章では MedLFRM を用いた離散関係属性付きネットワーク解析について述べる．とりわけ，この条件下でのリンク予測問題について取りあげる．

3.1 MedLFRM による離散関係属性付きネットワークの解析

ここまで我々が着目してきたモデル [8] では関係属性は連続値表現としてきたが，ソーシャルメディアなどのデータセットにおいて，関係属性はしばしば多次元離散表現とされている．したがって，これらのデータセットに対するリンク予測問題を考えるにあたって関係属性を多次元離散表現とする必要がある．

これを踏まえて，X_{ij} をこれまで $X_{ij} \in \mathbb{R}^D$，つまり各成分が $[0,1]$ の D 次元実数値ベクトルとしていたが，$X_{ij} \in \mathbb{I}^R$，つまり各成分が $\{0,1\}$ の R 次元二値ベクトルとする．以下では X_{ij} を以上のように離散関係属性であると再定義して議論する．

3.2 MedLFRM によるパラメータの推定

次に MedLFRM によるパラメータの推定方法について述べる．Zhu による手法 [8] に従った推定方法を以下に述べる．

打ち切り平均場近似 (truncated mean field approximation) [10] によって，$p(\boldsymbol{\nu}, Z, \Theta)$ を次のように表す．

$$p(\boldsymbol{\nu}, Z, \Theta) = p(\Theta) \prod_{k=1}^{K} p(\nu_k|\gamma_k) (\prod_{i=1}^{N} p(Z_{ik}|\psi_{ik})) \tag{9}$$

$p(\nu_k|\gamma_k)$ はベータ分布 $\mathrm{Beta}(\gamma_{k1}, \gamma_{k2})$ からサンプリングされたものであり，$p(Z_{ik}|\psi_{ik})$ は $\mathrm{Bernoulli}(\psi_{ik})$ からサンプリングされたもの，そして K は打ち切りレベル (注2) である．従って，MedLFRM の問題は以下の手順を交互に反復することで解くことができる．

（1）$p(\Theta)$ の推定

$p(\boldsymbol{\nu}, \mathbf{Z})$ が与えられた時，部分問題を以下の等価制約形で書くことができる．

$$\min_{p(\Theta),\boldsymbol{\xi}} \mathrm{KL}(p(\Theta)|p_0(\Theta)) + C \sum_{(i,j)\in\mathcal{E}} \xi_{ij} \tag{10}$$

$$\text{s.t.}: Y_{ij}(\mathrm{Tr}(\mathbb{E}[W]\bar{\mathbf{Z}}_{ij}) + \mathbb{E}[\eta]^{\mathsf{T}} X_{ij}) \geqq \ell - \xi_{ij},\ \forall (i,j) \in \mathcal{E}$$

$\bar{\mathbf{Z}}_{ij} = \mathbb{E}_p[Z_j^{\mathsf{T}} Z_i]$ はノード i，j の潜在特徴の内積の期待値であり，$\boldsymbol{\xi}$ はソフトマージンを実現するためのスラック変数である．ラグランジュ双対理論によって $p(\Theta)$ の最適解を得ることができる．$p(\Theta)$ は以下のように書くことができる．

$$p(\Theta) \propto p_0(\Theta) \exp\left\{ \sum_{(i,j)\in\mathcal{E}} \omega_{ij} Y_{ij} (\mathrm{Tr}(W\bar{\mathbf{Z}}_{ij}) + \eta^{\mathsf{T}} X_{ij}) \right\}$$

$\boldsymbol{\omega} = \{\omega_{ij}\}$ はラグランジュ乗数である．

(注2)：LFRM や MedLFRM は無限個の潜在特徴を仮定するが，近似のため十分大きな K のもと有限であると仮定する．

一般的に使用される標準正規事前分布 $p_0(\Theta)$ に対して，$p(\Theta)$ の最適解を得るとする．つまり以下のように書くことができる．

$$p(\Theta) = p(W)p(\eta) = \left(\prod_{kk'} \mathcal{N}(\Lambda_{kk'}, 1)\right)\left(\prod_{d} \mathcal{N}(\kappa_d, 1)\right)$$

上記の $\mathcal{N}(\Lambda_{kk'}, 1)$, $\mathcal{N}(\kappa_d, 1)$, それぞれの期待値は $\Lambda_{kk'} = \sum_{(i,j)\in\mathcal{E}} \omega_{ij} Y_{ij} \mathbb{E}[Z_{ik} Z_{jk'}]$, $\kappa_d = \sum_{(i,j)\in\mathcal{E}} \omega_{ij} Y_{ij} X_{ij}^d$ とする．このとき双対問題は以下のようになる．

$$\max_{\boldsymbol{\omega}} \sum_{(i,j)} \ell\omega_{ij} - \frac{1}{2}(\|\Lambda\|_2^2 + \|\kappa\|_2^2)$$
$$\text{s.t.} : 0 \leqq \omega_{ij} \leqq C,\ \forall (i,j) \in \mathcal{E}$$

よって，部分問題は以下のように書き直すことができる．

$$\min_{\Lambda,\kappa,\boldsymbol{\xi}} \frac{1}{2}(\|\Lambda\|_2^2 + \|\kappa\|_2^2) + C \sum_{(i,j)\in\mathcal{E}} \xi_{ij} \tag{11}$$
$$\text{s.t.} : Y_{ij}(\mathrm{Tr}(\Lambda \bar{\mathbf{Z}}_{ij}) + \kappa^{\mathsf{T}} X_{ij}) \leqq \ell - \xi_{ij},\ \forall (i,j) \in \mathcal{E}$$

これは SVM(Support Vector Machine) の二値分類問題の形式と一致している．従って，SVMlight(注3) や LibSVM(注4) などによって解くことができる．

（2）$p(\boldsymbol{\nu}, Z)$ の推定

$p(\Theta)$ が与えられると部分問題は以下のように書くことができる．

$$\min_{p(\boldsymbol{\nu},Z)} \mathrm{KL}(p(\boldsymbol{\nu}, Z)|p_0(\boldsymbol{\nu}, Z)) + C\mathcal{R}_\ell(p(Z, \Theta))$$

打ち切り平均場推定によって以下の式が得られる．

$$\mathrm{Tr}(\Lambda \bar{\mathbf{Z}}_{ij}) = \begin{cases} \psi_i \Lambda \psi_j^{\mathsf{T}} & \text{if } i \neq j \\ \psi_i \Lambda \psi_i^{\mathsf{T}} + \sum_k \Lambda_{kk} \psi_{ik}(1 - \psi_{ik}) & \text{if } i = j \end{cases}$$

マージンの制約は $\boldsymbol{\nu}$ によらないので，$p(\boldsymbol{\nu})$ は Doshi-Velez らと同じ解法を用いる [11].

$p(Z)$ は劣勾配法を用いることで解く．ここで観測されたリンクの組の集合 $\mathcal{E}$ を以下のようにおく．

$$\mathcal{E}_i = j : j \neq i,\ (i,j) \in \mathcal{E} \text{ and } Y_{ij} f(X_{ij}) \leqq \ell$$
$$\mathcal{E}'_i = j : j \neq i,\ (j,i) \in \mathcal{E} \text{ and } Y_{ji} f(X_{ji}) \leqq \ell$$

$g(x) \leqq \ell$ なら $\partial_x h_\ell(g(x))$ は $-\partial_x g(x)$ となり，そうでないなら 0 である．従って $\mathcal{R}_\ell$ の劣勾配は以下のようになる(注5).

$$\begin{aligned}\partial_{\psi_{ik}} \mathcal{R}_\ell = &- \sum_{j\in\mathcal{E}_i} Y_{ij} \Lambda_{k\cdot} \psi_j^T - \sum_{j\in\mathcal{E}'_i} Y_{ji} \psi_j \Lambda_{\cdot k} \\ &- \mathbb{I}(Y_{ii} f(X_{ii}) \leqq \ell) \\ &\quad Y_{ii}(\Lambda_{k\cdot} \psi_i^T + \psi_i \Lambda_{\cdot k} + \Lambda_{kk}(1 - 2\psi_{ik}))\end{aligned}$$

$\Lambda_{k\cdot}$ は Λ の k 番目の行であり，$\Lambda_{\cdot k}$ は Λ の k 番目の列である．また，$\mathbb{I}(\cdot)$ は指標関数である．部分問題の劣勾配を 0 とすることで，ψ_{ik} の更新式を得ることができる(注6).

$$\psi_{ik} = \Phi\left(\sum_{j=1}^{k} \mathbb{E}_p[\log \nu_j] - \mathcal{L}_k^\nu - C\partial_{\psi_{ik}} \mathcal{R}_\ell\right) \tag{12}$$

$\mathcal{L}_k^\nu$ は $\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)]$ の下界である．

4. 実　　験

この章では前述にある通り，実問題のデータセットを用いてリンク予測問題に対する実験を行い，その結果について考察する．また，実験に用いるハイパーパラメータを決定するための予備実験も行う．

4.1 データセット

実験には Facebook のネットワークデータセット(注7) を用いる．アカウントをノード，アカウント間の関係をリンクとする．このデータセットは全ノード数が 4039，全リンク数が 88234 のデータセットであるが，10 個のデータセットに分割されている．実験で用いたものはこの中の 3 つのデータセットであり，それぞれをデータセット A，B，C と表す．それぞれのデータセットに含まれるノード数やリンク数，属性数は Table 1 に示す．また，このデータセットは Facebook であるため，グラフ構造の隣接行列は対称であると言うことができる．

データセットをそれぞれランダムに 5 分割し，1 つを評価に用いるテストデータ，1 つを自由パラメータの決定に用いる検証データ，残った 3 つをモデルの潜在変数と未知パラメータの推定に用いる訓練データとする．

Table 1　A summary of the datasets used.

	no. of nodes	no. of edges	no. of observed features
data A	150	1693	105
data B	61	270	48
data C	52	146	42

4.2 実 験 設 定

次にハイパーパラメータについて述べる．打ち切りレベル K は全ての実験において $K = 50$ と設定した．また，各関係にリンクが存在する確率は，データセット A においては約 0.075，データセット B においては約 0.072，データセット C においては 0.054 であるため，いずれのデータセットもリンクの有無という点で不均衡であると言える．したがって正のデータに対しては C^+，負のデータに対しては C^- という異なった正則化定数を用いるとする．本稿の実験では全て $C^+ = 10C^- = 10C$ とし，$C = 1.0$ と設定した．

ハイパーパラメータ α と ℓ に関しては以下に述べる予備実験によって決定する．打ち切りレベル K と正則化定数 C に関しては上記の通りに，損失パラメータ ℓ は初期値として $\ell = 1$ と

(注3)：http://svmlight.joachims.org/

(注4)：http://www.csie.ntu.edu.tw/ cjlin/libsvm

(注5)：Zhu の論文 [8] には，右辺第 3 項が $-\mathbb{I}(Y_{ii} f(X_{ii}) \leqq \ell) Y_{ii}(\Lambda_{kk}(1 - \psi_{ik}) + \Lambda_{k\cdot} \psi_i^{\mathsf{T}})$ となっているが，正しくは本稿の本文のものである．

(注6)：Zhu の論文 [8] では，$\psi_{ik} = \Phi\left(\sum_{j=1}^k \mathbb{E}_p[\log \nu_j] - \mathcal{L}_k^\nu + C\partial_{\psi_{ik}} \mathcal{R}_\ell\right)$ となっているが，本稿の本文のものが正しい式である．

(注7)：http://snap.stanford.edu/data/egonets-Facebook.html

設定し，α を $\alpha \in \{0.125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0\}$ の範囲で値を変更させながら訓練データを用いてモデルのパラメータ推定を行う．その後，検証データに対して AUC(Area Under the ROC Curve) を記録する．これは二値分類問題における精度評価法の一つであり，ROC 曲線の積分値を示す．理想的な予測をした場合 1 を，最も悪い予測をした場合で 0.5 を記録する．AUC の記録をデータセット A，B，C で行い，各データセットの検証データにおける AUC の期待値を求め，最も精度のよいものを最適なハイパーパラメータ α の値とする．次に前述したとおりの K，C，そして上述の要領で決定したハイパーパラメータ α を用いて損失パラメータ ℓ を決定する．$\ell \in \{1, 5, 9, 13\}$ とし，この範囲で値を変更させながら訓練データを用いてモデルのパラメータ推定を行う．その後，α の決定方法と同様にして検証データに対して AUC を記録し，各データセットにおける AUC の期待値を求め，最も精度の良いものを最適な損失パラメータ ℓ とする．

モデルで推定するパラメータの初期化については以下の通りである．W は $[0, 0.1]$ の区間で一様に初期化，ψ は 0.5 に $[0, 0.001]$ の区間で一様に分布したランダムノイズを加えたものに初期化，そして η の平均は 0 と初期化する．

4.3 評価方法

予備実験の評価方法は，パラメータを変更させながら 4.1 節で述べた検証データに対する AUC を記録する．実験の評価方法は予備実験によって推定されたモデルを使用し，4.1 節で述べたテストデータに対する AUC を記録する．

また，ノードに与えられた属性を考慮した場合と考慮しない場合のモデルをそれぞれ推定し，結果を比較する．

4.4 予備実験結果と考察

4.4.1 結　　果

予備実験の結果である α の値を変更させた時の検証データに対する AUC を Fig. 2 に示す．各 AUC の値はデータセット A，B，C で記録された AUC の期待値である．α 以外のハイパーパラメータは 4.2 節で述べたとおり，$K = 50$，$C^+ = 10.0$，$C^- = 1.0$，$\ell = 1$ とする．グラフの縦軸は AUC の値，横軸は α の値を示している．また，赤いグラフが関係属性を考慮した場合のグラフであり，青いグラフが関係属性を考慮しない場合のグラフである．

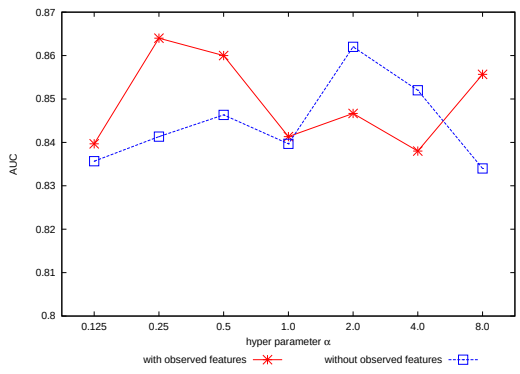

Fig. 2　AUC with varying hyper parameter α.

次に ℓ の値を変更させた時の検証データに対する AUC を Fig. 3 に示す．各 AUC の値はデータセット A，B，C で記録された AUC の期待値である．ℓ 以外のハイパーパラメータは α を除いて前述の通りとする．α は Fig. 2 から最も良い AUC が記録された値を使用する．つまり，関係属性を考慮する場合は $\alpha = 0.25$，考慮しない場合は $\alpha = 2.0$ とする．グラフの縦軸は AUC の値，横軸は ℓ の値を示している．また，赤いグラフが関係属性を考慮した場合のグラフであり，青いグラフが関係属性を考慮しない場合のグラフである．

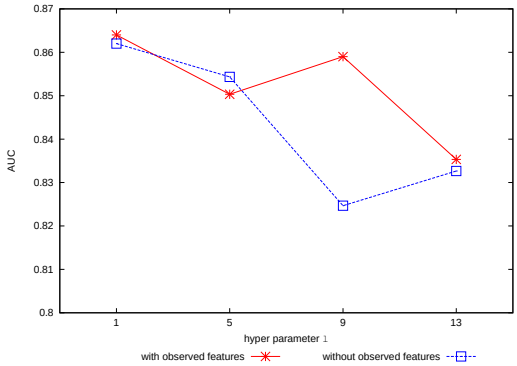

Fig. 3　AUC with varying hyper parameter ℓ.

4.4.2 考　　察

Fig. 2 から，ハイパーパラメータ α の値によってリンク予測精度が大きく変わることがわかる．さらにそれは関係属性を考慮する場合としない場合では振る舞いが異なることもグラフから見るができる．ここからハイパーパラメータ α の値が適切でないと予測精度に大きな差が出てしまう可能性があると考えることができる．

また，関係属性を考慮する場合は $\alpha = 0.25$ において，考慮しない場合は $\alpha = 2.0$ において最もリンク予測の精度がよいが，これらの値を比較すると関係属性を考慮する場合の方がリンク予測の精度がよい．このことから，ノードに与えられる関係属性を考慮するモデルの方がより精度の高いリンク予測が可能であることがわかる．

Fig. 3 からも同様に，ハイパーパラメータ ℓ の値によってリンク予測精度が大きく変わることがわかる．中でも関係属性を考慮しない場合において $\ell = 9$ の時に予測精度が大きく下がっている．ここからハイパーパラメータ ℓ の値が適切な値でないと予測精度が大きく落ちてしまう可能性があると考えることができる．特に関係属性を考慮しない場合はそれが顕著に表れると予想できる．

また，関係属性を考慮する場合と考慮しない場合，ともに $\ell = 1$ において最もリンク予測の精度がよいが，これらの値を比較すると関係属性を考慮する場合の方がわずかであるがリンク予測の精度がよい．このことから，ノードに与えられる関係属性を考慮するモデルの方がより精度の高いリンク予測が可能であることがわかる．

4.5 実験結果と考察

4.5.1 結　　果

実験結果であるリンク予測の結果を Table 2 に示す．これは予備実験において最適と判断したハイパーパラメータを用いて推定されたモデルによって各データセットのテストデータに対する AUC を記録したものである．予備実験からハイパーパラメータ α と ℓ は属性を考慮する場合は $\alpha = 0.25$，$\ell = 1$，考慮しない場合は $\alpha = 2.0$，$\ell = 1$ としたモデルを用いた．

Table 2　AUC of test data.

	with observed features	without observed features
data A	0.853	0.862
data B	0.912	0.887
data C	0.820	0.761
average	0.862	0.837

4.5.2 考　　察

Table 2 から，データセット A においては関係属性を考慮しないほうがリンク予測精度はわずかに高いが，データセット B とデータセット C，そして各データセットの平均値では関係属性を考慮したほうが予測精度が高いことが見て取ることができる．ここから予備実験と同様に，関係属性を考慮するモデルの方がより精度の高いリンク予測が可能であることがわかる．

データセット A において関係属性を考慮しないほうが AUC において高い値を記録している点に関してであるが，Table 1 で示しているようにデータセット A の各ノードには属性数が 105 個与えられている．ここから，ノードに与えられる属性数が多すぎるとリンク予測に不必要な属性が含まれる可能性が高くなるため，リンク予測の精度が属性を考慮しない場合と同等またはそれ以下となる場合があることがわかる．したがって，ノードに与えられる属性は詳細すぎないことが望ましいと言える．

5. おわりに

本稿では，ソーシャルメディアにおいてしばしばみられる離散関係属性付きのネットワークに対するリンク予測問題について，マージン最大化潜在特徴関係モデル (MedLFRM) を用いて評価を行った．

実問題である Facebook のデータセットを用いて，関係属性を考慮した場合と考慮しない場合においてそれぞれ実験し，テストデータに対する AUC を用いて比較した．その結果，関係属性を考慮した場合のほうがそうでない場合よりもリンク予測精度がよいことが確認された．また，ハイパーパラメータが適切な値でないとリンク予測精度が下がるため，予測精度を向上させるためには適切な値にしないといけないことも確かめられた．さらに，ノードに与える関係属性の数が多すぎると，かえってリンク予測精度が悪くなる可能性があることも示された．

今後の課題として，正則化定数 C の値を交差検定などの手法によって適切な値を決定し，リンク予測問題においてどのような精度が得られるかを評価することである．本稿の予備実験によってハイパーパラメータの値が予測精度に影響することが示されたため，より精緻に適切な値を決定する必要があると考えられる．さらに，各ノードに与えられる属性を要約したモデルによるリンク予測精度の比較を行うことが挙げられる．また，関係属性予測問題に関するモデルの推定方法の精緻化及び評価が挙げられる．そして実問題のデータセットに対して関係属性の予測精度を評価することである．これによってノード間の繋がりの内容を予測することができ，より詳細なネットワーク解析が可能となると考えられる．

謝　　辞

本研究の一部は，科学研究費補助金基盤研究 (B) (23300039) の援助による．

文　　献

[1] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, Vol. 58, No. 7, pp. 1019–1031, 2007.

[2] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 635–644. ACM, 2011.

[3] Kurt T Miller, Thomas L Griffiths, and Michael I Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, Vol. 22, pp. 1276–1284, 2009.

[4] Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, Vol. 20, pp. 737–744, 2007.

[5] Thomas L Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems*, Vol. 18, pp. 475–482, 2005.

[6] Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems*, Vol. 12, pp. 470–476, 1999.

[7] Tony Jebara. *Machine learning: discriminative and generative*. Springer, 2004.

[8] Jun Zhu. Max-margin nonparametric latent feature models for link prediction. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[9] Yee W Teh, Dilan Görür, and Zoubin Ghahramani. Stick-breaking construction for the indian buffet process. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pp. 556–563, 2007.

[10] Christopher M Bishop, et al. *Pattern recognition and machine learning*. springer New York, 2006.

[11] Finale Doshi, Kurt Tadayuki Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the indian buffet process. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 137–144, 2009.