Vol.2009-MPS-75 No.24
2009/9/11

時間変化に対応する無限関係モデルの提案

石黒勝彦[†1] 岩田具治[†1] 上田修功[†1]

近年，WWW や人間関係ネットワークなどオブジェクト間の関係を表す関係データの解析に大きな興味が寄せられている．関係データから構造を抽出するモデルとして，無限関係モデル (IRM) などが著名である．しかし，多くの既存のモデルは静的なデータを対象としたモデル故，関係が時間によって変化する，より現実的な関係データに対しては十分なモデルとは言えない．本論文では，IRM を拡張し，時間的に変化する関係データ解析のための新たなモデルを提案する．人工データおよび実データを用いた実験によりその有効性を確認する．

Dynamic Infinite Relational Model for Time-dependent Relational Data Analysis

Katsuhiko Ishiguro,[†1] Tomoharu Iwata [†1] and Naonori Ueda[†1]

Analysis of relational data such as the WWW and social networks' structures has drawn many attentions recently. For this purpose, many models including the infinite relational model (IRM) have been proposed. Since the most existing models are for static relations, it is insufficient for dynamic relational data analysis where relations vary with time. In this paper, we extend IRM to a dynamic model to solve this problem. We show the usefulness of the model through experiments with synthetic and real world data sets.

1. はじめに

現在，インターネット上の検索技術は，ホームページ間のリンク情報を用いたものが主流である．また，SNS の様に人と人のローカルなつながりによるネットワークに注目するサービスが多数提案されている．このように，近年ネットワーク構造に代表される" 関係データ" の重要性が認識されており，様々な研究が盛んになっている[1),2)]．

関係データモデリングの手法としては stochastic block model (SBM)[3)] とその拡張である inifinite relational model (IRM)[4)] が著名である．これらのモデルは，オブジェクト間の関係データから，そのオブジェクト集合を，クラスタ間の関係を最も適切に表現するクラスタに分割する．SBM では，クラスタリングの際，クラスタ数を事前に決定しておく必要があるが，IRM では，ノンパラメトリックベイズの枠組みでデータから最適なクラスタ数が自動的に推定される．

現実の関係データの中には，オブジェクト間の関係が時間的に変化するケースが数多く存在する．例えば，WWW ページのリンクなどは，ある時話題になったページにはリンクが集中するが，ブームが過ぎればそれらのリンクは廃れてしまう．また，ブームを追いかける集団 (ニュースサイトなど) は次のブームを追いかけて次々とリンク先を変化させるが，お互いの結合が強い SNS 上のコミュニティメンバは，流行に流されず，時間が経過しても高い確率でお互いをリンクするコミュニティに帰属し続ける可能性が高い．その一方で，コミュニティ自体が合併などによって消滅したり，新たに生成されることもある．

しかし，SBM, IRM 等の従来モデルは，時間変化を直接モデル化していないため，時間変化する関係データへの適用は困難である．時間発展する関係データのためのモデルも提案されているが[5),6)], これらの手法はクラスタ数を事前に決定しておく必要があるなど，総じて既存の手法は関係性のダイナミックスを捉えることができないという点で十分とは言えない．

本論文では，IRM を，時間変化する関係データのために拡張した動的無限関係モデル (dynamic IRM, dIRM) を新たに提案する．本モデルでは，ノンパラメトリックベイズに基づき，クラスタ数のダイナミックス，および，各オブジェクトの帰属確率のダイナミックスのための事前分布を導入することによりこれらの問題に対処している．

2. Infinite Relational Model(無限関係モデル)

N 個のオブジェクトからなるオブジェクト集合 $D = \{1, 2, \ldots, N\}$ 上の二値の二項関係 $X : D \times D \to \{0, 1\}$ を考える (**図 1**).

IRM は，N オブジェクト内で観測された関係データ $X = \{x_{i,j} \in \{0, 1\};\ 1 \le i, j \le N\}$ から，オブジェクト集合を複数のクラスタに分割する．クラスタ分割の際にノ

†1 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

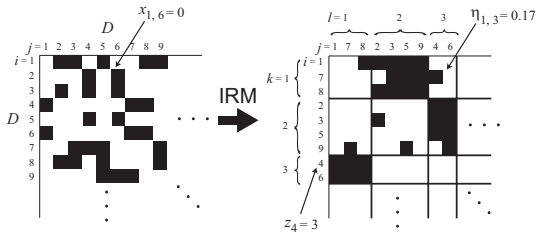

図 1 ブロックモデル (IRM) の例
Fig. 1 Example of block models (IRM).

ンパラメトリックベイズモデルの一種である Dirichlet Process Mixture(DPM)[7] モデルを用い，クラスタ数も同時に推定する点が IRM の特徴である．$x_{i,j} \in \{0,1\}$ は，オブジェクト i, j 間での関係の有無を表し，$x_{i,j} = 1(0)$ ならばオブジェクト i, j 間に関係が存在する (しない) ことを意味する．尚，この関係は有向，無向のどちらでも良く，無向ならば $x_{i,j} = x_{j,i}$ となる．

IRM は潜在クラスタを仮定する X の生成モデルで，以下の様に表わされる．

$$\boldsymbol{\beta}|\gamma \sim \text{Stick}(\gamma) \tag{1}$$
$$z_i|\boldsymbol{\beta} \sim \text{Multinomial}(\boldsymbol{\beta}) \tag{2}$$
$$\eta_{k,l}|\xi,\psi \sim \text{Beta}(\xi,\psi) \tag{3}$$
$$x_{i,j}|Z,H \sim \text{Bernoulli}\left(\eta_{z_i,z_j}\right) \tag{4}$$

式 (1) では，無限次元のクラスタ混合比ベクトル $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ を生成する．右辺の Stick(γ) とは，DPM モデルを用いる際に利用する stick-breaking process[8] を表す．具体的には，長さ 1 の棒を，v_1 対 $1-v_1$ の比で折り，v_1 に対応する部分の長さを β_1 とする．次いで，残された棒 ($1-v_1$ に対応する部分) をさらに v_2 対 $1-v_2$ の比で折り，v_2 に対応する部分の長さを β_2 とする．この操作を繰り返すことにより $\beta_k = v_k \prod_{l=1}^{k-1}(1-v_l)$ を得る．v_k はベータ分布 Beta$(1,\gamma)$ から生成する．続いて，混合比 $\boldsymbol{\beta}$ に従ってオブジェクト i が帰属するクラスタ $z_i = k, k = 1, 2, \ldots, \infty$ をサンプリングする (式 (2))．

次に，式 (3) に従い，クラスタ k, l 間の関係の強さを表すパラメータ $\eta_{k,l}$ をサンプリングする．この値は，図 1 において，(k,l) で表されるブロック内の $x_{i,j}$ が 1 となる確率を表す．式 (4) では，$Z = \{z_i\}_{i=1}^N$，$H = \{\eta_{k,l}\}_{k,l=1}^{\infty}$ が与えられたときに観測量 $x_{i,j}$ をベルヌーイ分布から生成する．以上の過程の変数間の依存関係を表すグラフィカルモデルを**図 2**(a) に示す．

3. 動的 IRM モデル

3.1 対象とする関係データ

本論文で対象とする時刻データを含む関係データは $X = \{x_{t,i,j} \in \{0,1\};\ i, j = 1, 2, \ldots, N,\ t = 1, 2, \ldots, T\}$ で表わされるものとする．ここで，$x_{t,i,j} = 1(0)$ は時刻 t においてオブジェクト i, j に関係がある (ない) ことを意味する．時刻 t は離散時間とし，T 時刻までのデータが観測されているとする．また，異なる時刻のオブジェクト間では関係は定義されないものとする．

我々は，時系列関係データのモデル化では以下の 3 つの性質をモデル化することが必要であると考えた．

[1] クラスタリング結果が隣接時刻間で高い相関を有する事．
[2] クラスタリングの時間発展は一様でない事．
[3] クラスタ数の時間変化を許容する事．

例えば，インターネット上のリンク関係が時間と共に変化することから各オブジェクトも時間と共に異なるクラスタ間を遷移すると予想される．この際，隣接時刻間であるオブジェクトが同じクラスタに帰属しやすいという性質 [1] の仮定は妥当である．また，組織内の人間関係を大きく変化させる部署の合併や分裂，あるいはウェブ上での話題の流行は突発的に起こることから性質 [2] も自然な要請である．またクラスタの生成や消滅，合併・分裂が起きる以上，性質 [3] は必要な性質である．

3.2 ナイーブな拡張

時間的に変化する関係データに対処するための IRM の幾つかのナイーブな拡張法について検討する．

最も単純には，時刻データを含んだ関係データ X から時刻データを含まない関係データ $\tilde{X}$ を生成して通常の IRM を適用する方法が考えられる．例えば

$$\tilde{x}_{i,j} = \begin{cases} 1 & \frac{\sum_{t=1}^{T} x_{t,i,j}}{T} > \sigma \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

とすれば良い．ここで，σ は閾値である．この $\tilde{X}$ に対して IRM を適用し，そのクラスタリング結果を全時刻でのオブジェクトクラスタリングに適用する．明らかにこの方法では，関係性の時間変化は無視され，本論文の目的に整合しない．

クラスタリングの時間変化をモデル化する方法として，以下の様に，オブジェクト i のクラスタ帰属変数 z_i を時刻 t 依存 ($z_{t,i}$) とし，かつ，全時刻でその生成分布パラメータ $\boldsymbol{\beta}$ を共有化するモデルも考えられる (図 2(b))．なお，$Z_t = \{z_{t,i}\}_{i=1}^N$，$H = \{\eta_{k,l}\}_{k,l=1}^{\infty}$ を表す．

$$\boldsymbol{\beta}|\gamma \sim \text{Stick}(\gamma) \tag{6}$$
$$z_{t,i}|\boldsymbol{\beta} \sim \text{Multinomial}(\boldsymbol{\beta}) \tag{7}$$
$$\eta_{k,l}|\xi,\psi \sim \text{Beta}(\xi,\psi) \tag{8}$$
$$x_{t,i,j}|Z_t,H \sim \text{Bernoulli}\left(\eta_{z_{t,i},z_{t,j}}\right) \tag{9}$$

しかし，このモデルでは全ての t, i についてクラスタインデックス $z_{t,i}$ が $\boldsymbol{\beta}$ 所与の下で条件付き独立となり，時刻 $t-1$ と時刻 t でのクラスタリングに直接の依存関係がモデル化されていない．換言すれば，時刻の順序が無視されたモデル故，時間発展のモデル化としては適切ではない．

3.3 提案モデル

前節までの考察に基づき，先にあげた 3 つの性質を満たすべく，IRM を以下のように拡張する．同時に対応するグラフィカルモデルを図 2(c) に示す．なお，式

中で $\Pi = \{\boldsymbol{\pi}_{t,k} : t = 1,\ldots,T, k = 1,\ldots,\infty\}$ である．

$$\boldsymbol{\beta}|\gamma \sim \text{Stick}(\gamma) \tag{10}$$

$$\boldsymbol{\pi}_{t,k}|\alpha_0,\kappa,\boldsymbol{\beta} \sim \text{DP}\left(\alpha_0+\kappa, \frac{\alpha_0\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k}{\alpha_0+\kappa}\right) \tag{11}$$

$$z_{t,i}|z_{t-1,i},\Pi_t \sim \text{Multinomial}\left(\boldsymbol{\pi}_{t,z_{t-1,i}}\right) \tag{12}$$

$$\eta_{k,l}|\xi,\psi \sim \text{Beta}(\xi,\psi) \tag{13}$$

$$x_{t,i,j}|Z_t,H \sim \text{Bernoulli}\left(\eta_{z_{t,i},z_{t,j}}\right) \tag{14}$$

$\boldsymbol{\pi}_{t,k} = (\pi_{t,k,1},\pi_{t,k,2},\ldots,\pi_{t,k,l},\ldots)$ は，時刻 $t-1$ においてクラスタ k に帰属していたオブジェクトが，時刻 t で第 l クラスタに遷移する ($l = 1,2,\ldots$) 確率を表す．$\boldsymbol{\pi}_{t,k}$ は無限次元のディリクレ分布 ($\pi_{t,k,l} > 0$ かつ $\sum_{l=1}^{\infty}\pi_{t,k,l} = 1$) に相当し，ノンパラメトリックベイズで多用されるディリクレ過程 (Dirichlet Process:DP) を用いて生成する．

DP は基底分布 G_0 と正のパラメータ α_0 を用いて定義される，確率分布に対する分布である．分布 G が DP に従う時，$G \sim \text{DP}(\alpha_0, G_0)$ と表記する．G は確率分布故，新たな確率変数 θ が G から生成されるとすると，$\theta_1,\ldots,\theta_{i-1}$ が生成された時，DP では以下のような確率で θ_i が生成される．

$$P(\theta_i|\theta_{1:i-1}) = \frac{\alpha_0}{i-1+\alpha_0}G_0(\theta_i) + \sum_{k=1}^{K}\frac{m_k}{i-1+\alpha_0}\delta_{\theta_{(k)}}(\theta_i)$$

$\delta_x(y)$ は $x = y$ の時 1，それ以外は 0 となるデルタ関数を表す．また，$\theta_1,\ldots,\theta_{i-1}$ の種類が K(異なり数が K) のとき $\theta_{(k)}$ は k 種類目のパラメータであり，$m_{(k)}$ は $i-1$ 個の θ のうち $\theta_{(k)}$ と同じ値を取ったものの総数である．上式は，θ_i の値がこれまでに生成されたパラメータから出現回数に比例して選択されること，さらに一定の割合で新しいパラメータを基底分布 (θ の事前分布) から生成することを示している．

翻って，式 (12) の場合，α_0 が $\alpha_0+\kappa$ に，G_0 が $(\alpha_0\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k)/(\alpha_0+\kappa)$ に相当する．$\boldsymbol{\delta}_k$ は第 k 要素の値が 1，他は全て 0 のベクトルである．つまり，$(\alpha_0\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k)$ の第 $j(\neq k)$ 要素は $\alpha_0\beta_j$，第 k 要素は $\alpha_0\beta_k+\kappa$ となり、k 番目の要素にバイアスが付与される．$(\alpha_0\boldsymbol{\beta}+\kappa\boldsymbol{\delta}_k)/(\alpha_0+\kappa)$ を DP の基底分布とすることで，$\boldsymbol{\pi}_{t,k}$ の第 k 要素 $\pi_{t,k,k}$ の値が $\pi_{t,k,l}$ $(l \neq k)$ に比べ $\kappa/(\alpha_0+\kappa)$ だけ確率値を加算させることになる．これにより，式 (12) での $z_{t,i}$ のサンプリングにおいて，$z_{t-1,i} = k$ の時，$z_{t,i} = k$ とサンプリングされる確率を相対的に大きくすることができ，性質 [1] を反映したモデル化が実現できる．

またクラスタ遷移確率のパラメータである $\boldsymbol{\pi}_{t,k}$ は時刻 t ごとにサンプリングされているため，性質 [2] を満たしている．さらに，提案モデルは IRM を踏襲しているので，性質 [3] を満たすことも明らかである．以上説明した提案モデルを，動的 IRM(dynamic IRM: dIRM) と呼ぶこととする．

dIRM は，Teh[9] および Fox[10] によって提案された infinite HMM モデルと関係が深い．これらのモデルとの最大の相違点は，9)，10) ではクラスタ間の遷移確率が時間不変であるが，dIRM モデルは遷移確率 $\boldsymbol{\pi}_{t,k}$ が時刻 t にも依存する点である．従って，ブームによるクラスタの人気の変化や突発的なクラスタの合併や分裂などが考えられる場合，時刻によって遷移確率が変化する dIRM のようなモデル化が妥当と言える．

なお，関係データに対する時系列モデルは，提案モデルの他にも 5)，6) などが最近提案されているが，いずれの手法もクラスタ数を事前に決定しておく必要があり，この点で我々の目的を満たさない．

3.4 学　　習

関係データ $X = \{x_{t,i,j}\}$ からパラメータ $z_{t,i}, \boldsymbol{\pi}_{t,k}, \eta_{k,l}, \boldsymbol{\beta}$ およびハイパーパラメータ $\gamma,\kappa,\alpha_0,\xi,\psi$ を学習する方法について説明する．本論文では beam サンプリング法に基づいて学習アルゴリズムを導出する．

beam サンプリングでは，補助変数 $U = \{u_{t,i}\}$ を導入して，Z などと同時にサンプリングする．この事で，無限個存在し得るクラスタ数を，U の条件付きの下で有限個に制限して学習を進めることが可能になる．

3.4.1 $u_{t,i}$ のサンプリング

$u_{t,i}$ の事前分布は一様分布と仮定する．また，$\eta,\boldsymbol{\pi}$ が既知の時，u,z と x の同時分布を次のように定義する．

$$p\left(x_{t,i,j},u_{t,i},u_{t,j},z_{t-1:t,i},z_{t-1:t,j}\right) = \mathbb{I}\left(u_{t,i} < \pi_{t,z_{t-1,i},z_{t,i}}\right) \mathbb{I}\left(u_{t,j} < \pi_{t,z_{t-1,j},z_{t,j}}\right) x_{t,i,j}^{\eta_{z_{t,i},z_{t,j}}}\left(1-x_{t,i,j}\right)^{1-\eta_{z_{t,i},z_{t,j}}}$$

ここで，$\mathbb{I}$ は，続く条件式が満たされれば 1，そうでなければ 0 の値をとる．この式から，U の事後分布が次のように求まる．

$$p(U|X,Z,\Pi,H,\boldsymbol{\beta}) = \prod_t\prod_i \frac{\mathbb{I}\left(u_{t,i} < \pi_{t,z_{t-1,i},z_{t,i}}\right)}{\pi_{t,z_{t-1,i},z_{t,i}}}$$

従って，各 $u_{t,i}$ は次の分布からサンプリングすれば良い．

$$u_{t,i} \sim \text{Uniform}\left(0,\pi_{t,z_{t-1,i},z_{t,i}}\right) \tag{15}$$

他の変数の事後分布も同様にして求めることが出来る．

3.4.2 $z_{t,i}$ のサンプリング

U の導入により，HMM における forward-backward アルゴリズムと類似のアルゴリズムで効率的に $z_{t,i}$ をサンプリングできる．まず HMM の場合と同様に，次のメッセージ変数を定義する．

$$p_{t,i,k} = p(z_{t,i} = k|X_{1:t},U_{1:t},\Pi,H,\boldsymbol{\beta}) \tag{16}$$

$z_{t,i}$ のサンプリングでは，まず上記のメッセージ変数を $t = 1$ から $t = T$ まで計算する．次に，この変数を用いて，$z_{t,i} = k$ の値を $t = T$ から $t = 1$ までサンプリングする，という二段階のステップを取る．

まず次の式を $t = 1$ から $t = T$ まで計算する．

$$p_{t,i,k} \propto p(x_{t,i,i}|z_{t,i} = k,H)\prod_{j\neq i}p\left(x_{t,i,j}|z_{t,i} = k,H\right) p\left(x_{t,j,i}|z_{t,i} = k,H\right)\sum_{l:u_{t,i}<\pi_{t,l,k}}p_{t-1,i,l} \tag{17}$$

次に，$t = T$ から $t = 1$ まで以下の式を計算して $z_{t,i}$ をサンプリングする．

$$p(z_{t,i} = k|z_{t+1,i} = l) \propto \pi_{t+1,k,l}p_{t,i,k}\mathbb{I}(u_{t+1,i} < \pi_{t+1,k,l}) \tag{18}$$

$\mathbb{I}(u_{t+1,i})$ によって，取りうる k の値が有限個に制限されるため，Z の値は有限 (K) 種しか持たない．

3.4.3 $\boldsymbol{\pi}_{t,k}$ のサンプリング

時刻 t において，$z_{t-1,i} = k$ かつ $z_{t,i} = l$ となるオブ

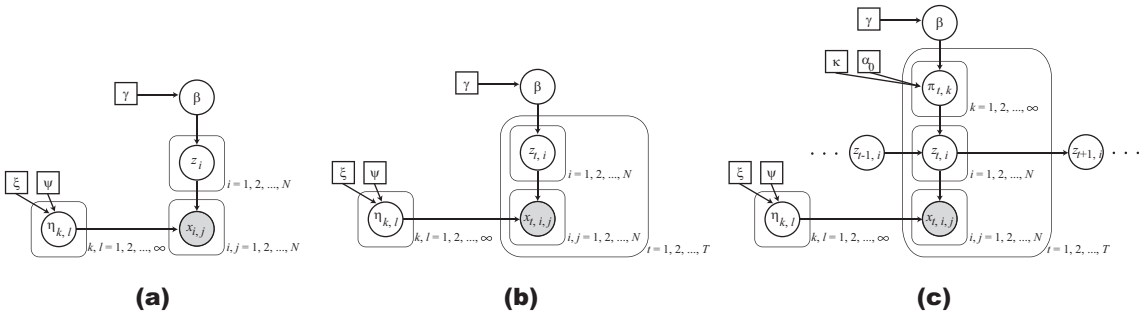

図 2 グラフィカルモデル．円ノードは変数，矩形ノードは定数を表す．影付きのノードは観測量を表す．(a)IRM モデル (Eqs.1–4)．(b)"tIRM" モデル (Eqs.6–9)．(c)dIRM モデル (Eqs.10–14).

Fig. 2 Graphical models of existing and proposed models. Circle nodes are the variables, and the square nodes are constants. Shaded nodes indicate the observations. (a)IRM (Eqs.1–4). (b)"tIRM" (Eqs.6–9). (c)dIRM(Eqs.10–14).

ジェクト i の数を $m_{t,k,l}$ とする．U,Z が所与の元で $\boldsymbol{\pi}_{t,k}$ は以下のように有限次元の事後分布 (Dirichlet) からサンプリングされる．ここで，$\beta_u = 1 - \sum_{k=1}^{K} \beta_k$ である．

$$\boldsymbol{\pi}_{t,k} \sim \mathrm{Dirichlet}(\alpha_0\beta_1 + m_{t,k,1}, \ldots, \alpha_0\beta_k + m_{t,k,k} + \kappa, \ldots, \alpha_0\beta_K + m_{t,k,K}, \alpha_0\beta_u) \quad (19)$$

3.4.4 $\eta_{k,l}$ のサンプリング

$z_{t,i} = k$, $z_{t,j} = l$ となる $x_{t,i,j}$ の数を全時刻に渡って加えたものを $N_{k,l}$，そのうち $x_{t,i,j} = 1$ となった観測値の数を $n_{k,l}$ とする．この時，グループ間の強さパラメータ $\eta_{k,l}$ は以下の事後分布からサンプリングされる．

$$\eta_{k,l} \sim \mathrm{Beta}\left(\xi + n_{k,l}, \psi + N_{k,l} - n_{k,l}\right) \quad (20)$$

3.4.5 β のサンプリング

U,Z が所与の下ではクラスタ数は有限 K 個なので，$\boldsymbol{\beta}$ の事後分布も通常の Dirichlet 分布となる．

$$\boldsymbol{\beta} \sim \mathrm{Dirichlet}\left(\sum_{t,k} \hat{R}_{t,k,1}, \sum_{t,k} \hat{R}_{t,k,2}, \ldots, \sum_{t,k} \hat{R}_{t,k,K}, \gamma\right) \quad (21)$$

$$\hat{R}_{t,k,l} = R_{t,k,l} - \delta_k(l) O_{t,k} \quad (22)$$

$R_{t,k,l}$ および $O_{t,k}$ は各々下記の式に従ってサンプリングする[10]．ここで，$s()$ は unsigned stirling number of the first kind とよばれる関数である．

$$p\left(R_{t,k,l} = r | z_{t,i}, \theta\right) = s\left(m_{t,k,l}, r\right) \left(\alpha_0\beta_l + \kappa\delta_k(l)\right)^r \frac{\Gamma\left(\alpha_0\beta_l + \kappa\delta_k(l)\right)}{\Gamma\left(\alpha_0\beta_l + \kappa\delta_k(l) + m_{t,k,l}\right)} \quad (23)$$

$$O_{t,k} \sim \mathrm{Binomial}\left(R_{t,k,k}, \frac{\kappa}{\kappa + \alpha_0\beta_k}\right) \quad (24)$$

3.4.6 ハイパーパラメータのサンプリング

ハイパーパラメータ $\gamma, \kappa, \alpha_0, \xi, \psi$ も同様にサンプリングにより同時推定可能であるが，具体的な計算式については省略する．

4. 実　　験

4.1 人工データによる評価

クラスタリングの正解が既知の人工データ 2 種を用いてモデルの定量的な評価を行った．一つ目の人工データ (synth1) は時間ステップ数 $T = 5$，オブジェクト数 $N = 16$，クラスタ数は $K = 4$ とした．一部のオブジェクト (延べ 6%) は時間に応じてクラスタ間を遷移する．二つ目の人工データ (synth2) は時間ステップ数 $T = 10$，オブジェクト数 $N = 54$，クラスタ数は $K = 6$ とした．このデータは，クラスタの一部が消滅もしくは新たなクラスタが発生する上に，オブジェクトのクラスタ間遷移もより頻繁に起こる様設定した (延べ 15%, **図 3**)．クラスタ間の関係の強さを表す $\eta_{k,l}$ は positive なクラスタ間では $\eta = 0.9$，negative なクラスタ間では $\eta = 0.1$(synth1) あるいは $\eta = 0.05$(synth2) の 2 種類のいずれかを選択した．与えられた $z_{t,i}$, $\eta_{k,l}$ に従って各時刻の観測量 $x_{t,i,j}$ を生成した．

実験では，上記の手続きに従って生成された観測データ X を用いて以下の 3 モデルを比較評価した．

(1) dIRM
(2) tIRM (時間インデックスを導入した IRM)
(3) 通常の IRM

tIRM は，式 (6)，式 (7)，式 (8)，式 (9) の生成モデルで定義される時間インデックスを導入した IRM である．ただし，dIRM と違い時間方向の依存関係がモデル化されていない．通常の IRM では，式 (5) において，$\sigma = 0.5$ としてクラスタリングした結果 z_i を各時刻でのオブジェクト i のクラスタリング $z_{t,i}$ と見なした．

本実験では Rand index[11] を利用した定量的評価を行った．Rand index とは，あるデータに対して 2 つのクラスタリング結果が与えられた時，2 つのクラスタリング結果の類似度を測る指標である．Rand index は非負であり，2 つのクラスタリング結果が完全一致した時に最大値 1 をとる．本実験では，各時刻 t においてクラスタリング推定結果 Z_t と正しいクラスタリング結果 $\hat{Z}_t$ の間で Rand index を計算し、T 時刻に渡る Rand index の算術平均をモデルの評価値とした．

各モデルによる Rand index の計算結果を**表 1** に示す．表より明らかなように，クラスタリングの時間変化を考慮しない IRM モデルや，tIRM の様に時間ステップ間の依存関係をしないモデルに比べ，提案した dIRM はより良く時間変化する関係データをモデル化

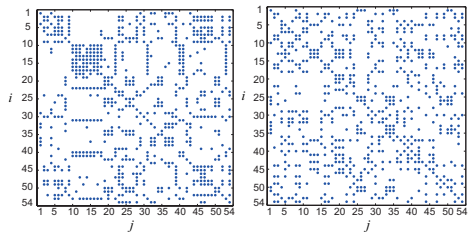

図 3 人工データ 2(synth2) の例．左: $t = 3$，右: $t = 8$．
Fig. 3 Examples of the synthetic data 2. Left: observations at $t = 3$. Right: observations at $t = 8$.

表 1 Rand index の計算結果
Table 1 Computed Rand indices for dIRM and other models.

Data	IRM	tIRM	dIRM (proposed)
synth1	0.7957	0.9462	**0.9819**
synth2	0.4331	0.7344	**0.8471**

できることが確認された．

4.2 実データによる評価

Enron e-mail dataset[12] を用いて実データ実験を行った．このデータは，Enron 社内のメールの一部を収集したもので，多くの研究で利用されている[5),13)]．

本実験では，$N = 151$ 人の Enron 社員の 2001 年 1 月から 12 月のメールデータを用いた．データを月ごと ($t = 1, 2, \ldots, 12$) に分離し，ある月 t に社員 i から社員 j に対して 1 通でもメールが発信されていれば $x_{t,i,j} = 1$，メールがなければ $x_{t,i,j} = 0$ とした．

Enron 事件については，いくつかの重要な時間軸が存在する．まず，2001 年 6 月 ($t = 6$) に Enron は 500 億ドルの売上を報告したが，利益率は低さなどから経営を不安視された．8 月 ($t = 8$) には株価急落を理由に CEO が辞任した．10 月に不正会計疑惑が報じられ，12 月 ($t = 12$) に Enron は倒産した．

このデータに対し dIRM を適用・学習した結果を説明する．まず，**図 4**(a) は学習されたクラスタ間の関係の強さ $\eta_{k,l}$ である．縦軸がメールの送信元，横軸がメールの受信元を表わす．次に，各クラスタに帰属した延べオブジェクト数を図 4(b) に示す．

これらの図より，クラスタ 4 は，他のクラスタとの関係が薄いことがわかる．つまり，そのクラスタ内のメンバだけでメールの関係を構築しているコミュニティであると考えられる．このクラスには，主に規制業種 (エネルギー関連）やガス，パイプライン部門の Vice President などが帰属している．

クラスタ 5 は Trader や経理関係など，金融・財務部門のメンバが多く集まっている．クラスタ 6 は”non active” なクラスタ，つまりほとんどメールのやりとりをしていないメンバが集まっている．クラスタ 7 は主に上位役員たちが集中している．クラスタ 9 には全時刻を通して延べ 3 人 (オブジェクト) だけが帰属している．但し，これらのオブジェクトは，5 月，8 月，10 月 ($t = 5, t = 8, t = 10$) という，Enron 事件において重要な月に特異的に多くのメンバにメッセージを発信している．5 月にクラスタ 9 に帰属したのは Enron America の CEO，8 月は Enron の Founder，そして 10 月は COO である．

クラスタ 9 の解析でも見たように，dIRM の一つの利点は，時刻ごとのクラスタ関係を追跡することが可能な点にある．図 4(c) は各クラスタへの帰属オブジェクト数を時刻ごとにプロットした図である．

この図からは興味深い点が幾つか散見される．まず，“non active” なクラスタ 6 は時間の経過と共に小さくなっている．これは，Enron がニュースを騒がせると共に社内の連絡も活性化したのであろう．次にクラスタ 5 であるが，このグループの構成員数は他のクラスタと比較して安定していることから，経理・金融関連の社員の結びつきが強固であることが伺える．

また，クラスタ 3 は 5 月から 7 月にのみ多くのメンバーを集めている．このクラスタのメンバーには幾人からの管理職や Vice President，Trader たちが含まれていること，さらにクラスタ 9 との関連が非常に強いことから，8 月の CEO 辞任の引き金となった株価下落などに関連している可能性がある．

以上の考察の信憑性は保証できないが，dIRM のクラスタリング結果から，上記のような直感的に妥当な関係性の時間的推移を分析することが可能となる．

5. ま と め

本論文では，関係データの解析モデルの一つである IRM を時間発展する関係データに適用できるよう拡張した動的 IRM(dynamic IRM，dIRM) モデルを提案した．dIRM モデルはクラスタのサイズや各オブジェクトの帰属クラスタがダイナミックに変化する時系列関係データを自然にモデル化，解析可能である．本論文では dIRM モデルの確率的生成モデルを説明し，その学習方法の一つとして beam サンプリングを利用することを提案した．人工データおよび実データを用いた実験で，良い性能を示すことを確認した．

謝辞 議論に参加していただいた，NTT の山田武士，持橋大地両氏に感謝する．

参 考 文 献

1) Liben-Nowell, D. and Kleinberg, J.: The Link Prediction Problem for Social Networks, *Proc. the 12th Int. Conf. Information and Knowledge Management*, pp.556–559 (2003).
2) 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).
3) Nowicki, K. and Snijders, T. A. B.: Estimation and Prediction for Stochastic Blockstructures, *J. Am.*

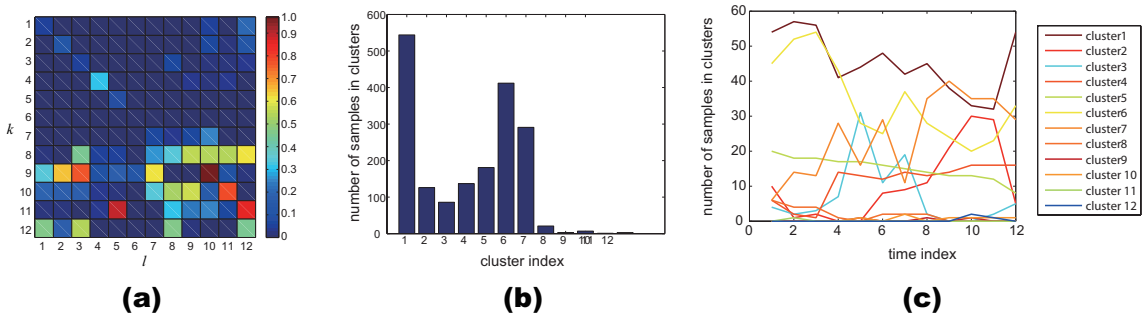

図 4 実データでの実験結果．(a) クラスタ間の関係パラメータ $\eta_{k,l}$ の学習結果．(b) 各クラスタに帰属したオブジェクト総数．(c) 各時刻において各クラスタに帰属したオブジェクト数．

Fig. 4 Experimental results on the real dataset. (a)Estimated $\eta_{k,l}$. (b)Total number of items belong to the clusters. (c)Number of items belong to the clusters at each time steps.

Stat. Assoc., Vol.96, No.455, pp.1077–1087 (2001).

4) Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. and Ueda, N.: Learning Systems Of Concepts With An Infinite Relational Model, *Proc. AAAI* (2006).
5) Fu, W., Song, L. and Xing, E.P.: Dynamic Mixed Membership Blockmodel for Evolving Networks, *Proc. ICML* (2009).
6) Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R.: A Bayesian Approach Toward Finding Communities and Their Evolutions in Dynamic Social Networks, *Proc. SIAM Int. Conf. Data Mining* (2009).
7) Ferguson, T.S.: A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, Vol.1, No.2, pp.353–355 (1873).
8) Sethuraman, J.: A Constructive Definition of Dirichlet Process, *Statistica Sinica*, Vol.4, pp.639–650 (1994).
9) Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet Process, *J. Am. Stat. Assoc.*, Vol.101, No.476, pp.1566–1581 (2006).
10) Fox, E., Sudderth, E., Jordan, M. and Willsky, A.: An HDP-HMM for Systems with State Persistence, *Proc. ICML* (2008).
11) Hubert, L. and Arabie, P.: Comparing partitions, *Journal of Classification*, Vol.2, No.1, pp.193–218 (1985).
12) CALO Project (A Cognitive Assistant that Learns and Organizes): Enron Email Dataset (2005).
13) Shetty, J. and Adibi, J.: Discovering Important Nodes through Graph Entropy: The Case of Enron Email Database, *Proc. the 3rd Int. Workshop on Link Discovery*, pp.74–81 (2005).