DEIM Forum 2018 F5-5

多様なノイズに頑健な関係データのクラスタリング

武田　悠佑† 　岩田　具治†† 　澤田　宏††

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916-5
†† 日本電信電話株式会社　NTT コミュニケーション科学基礎研究所
〒 619–0237 京都府相楽郡精華町光台 2-4
E-mail: †takeda.yusuke.tr0@is.naist.jp, ††{iwata.tomoharu,sawada.hiroshi}@lab.ntt.co.jp

あらまし　SNS におけるユーザネットワークから同じ嗜好を持つユーザのコミュニティを発見することや，EC サイトにおける購買履歴データから購入する商品の傾向が似ている顧客グループを抽出するといった応用が可能なことから，関係データのクラスタリングは重要視されている．一方で，既存のクラスタリング手法では実世界の関係データにおける多様なノイズが十分に考慮されておらず，データに多様なノイズが含まれている場合クラスタが適切に抽出できないという問題がある．そこで，関係データにおけるノイズを分離しつつクラスタリングを行うための確率モデルを提案する．人工データと実世界データを用いた実験により提案法の有効性を示す．

キーワード　確率的ブロックモデル, 関係データ, ノンパラメトリックベイズ，クラスタリング

1. はじめに

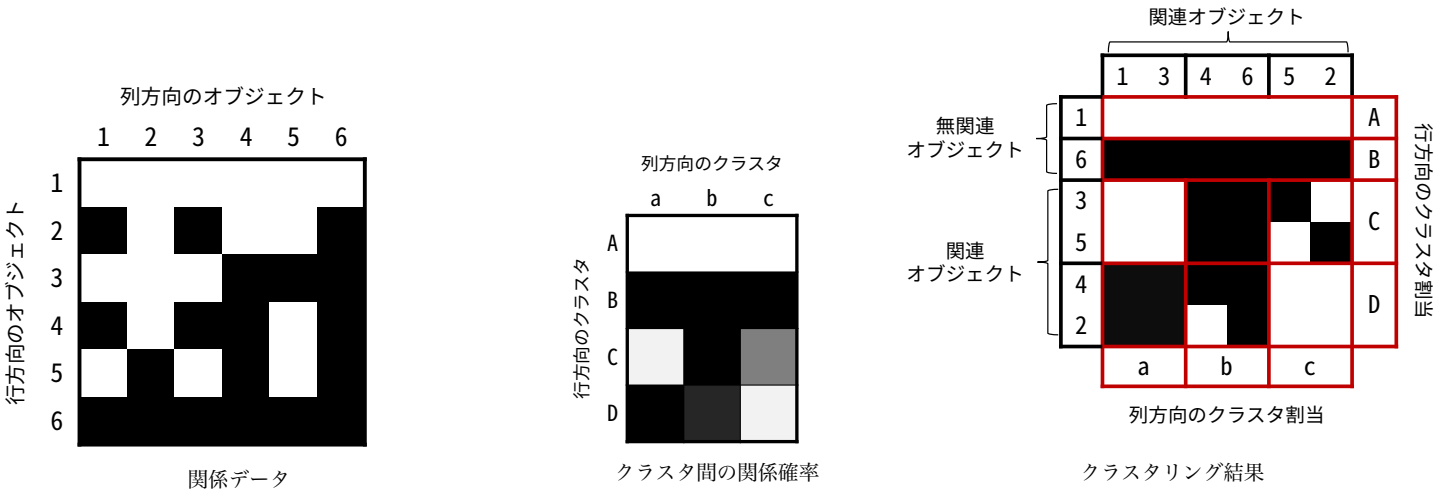
近年，関係データの解析に注目が集まっている．関係データとは，誰と誰が友達であるかといった SNS におけるユーザネットワークや，誰が何を買ったかといった EC サイトにおける購買履歴データに代表される,「ヒトとヒト」や「ヒトとモノ」といったオブジェクト間の関係を表現したデータの総称である．関係データでは個々のオブジェクトの性質が他のオブジェクトとの関係によって表現される．関係データを構成するオブジェクト集合を似た性質を持つオブジェクトのグループに分割することは，オブジェクトの性質やデータ全体の概要の把握において重要である．例えば，SNS におけるユーザネットワークから仲のいいユーザ同士のグループをコミュニティとして発見することは，同じコミュニティのユーザは嗜好が似ていると仮定すると，各ユーザの性質の推定に大きく寄与すると考えられる．また EC サイトにおける購買履歴データから購入する商品の傾向が似ている顧客グループを抽出することは，どのような商品がどのような顧客によって購入されているかの分析に役立つと考えられる．似た性質を持つオブジェクトのグループはクラスタ，オブジェクト集合を複数のクラスタへと分割することはクラスタリングと呼ばれ，関係データにおいて多くのクラスタリング手法が提案されている．

関係データのクラスタリング手法として確率的生成モデルである確率的ブロックモデル (SBM) [1] がよく知られている．SBM では関係データに潜在的な構造を仮定し，その推定を通してクラスタリングを行う．具体的には，個々のオブジェクト間の関係の有無はそれぞれが所属するクラスタによって決定されるという仮定を置いた上で，関係データ全体をよく表現するように各オブジェクトをクラスタリングする．SBM ではクラスタリングの過程でデータの潜在的な構造を推定するため，SNS のユーザネットワークから友達のグループを見つけるといったコミュニティ抽出のタスクに加えて，購買履歴データからあるユーザがある商品を買うかどうかを予測するといったリンク予測のタスクに対しても用いることができる [2] [3]．リンク予測やコミュニティ抽出における SBM や SBM をクラスタ数を自動推定するように拡張した無限関係モデル (IRM) [4] の有用性は多くの研究で示されている [5] [6].

一方で，SBM や IRM は関係データに含まれるノイズが考慮されていないという問題がある．ここでのノイズとは関係データ中で構造的な特徴を有さないオブジェクトである．具体的には，どの文書にどの単語が出現するかを表現した関係データの場合，ストップワードと呼ばれるようなほとんどの文書に出現する単語や，文書ジャンルによらずランダムに出現する記号文字といった単語などのように．他のオブジェクトとの関係の有無が相手のクラスタに依存しないようなオブジェクトである．SBM や IRM では，このような無関連なオブジェクトでもなんらかのクラスタへと割り当てるため，クラスタ構造が不明瞭になりクラスタが適切に推定できなくなる可能性がある．本研究ではこのようなオブジェクトを無関連オブジェクト，また無関連オブジェクトでないオブジェクトを関連オブジェクトと呼称する．

この問題に対して，IRM の拡張手法であるサブセット無限関係モデル (SIRM) [7] では各オブジェクトが関連オブジェクトか無関連オブジェクトかを判定する機構を導入し，無関連オブジェクトと判定されたオブジェクトについてはクラスタリングの対象から外すことで頑健なクラスタリングを試みている．しかし，全ての無関連オブジェクトが同一の確率で関係を持つと仮定するため，性質の異なる無関連オブジェクトに対応できない．

一方で，先述した単語と文書の関係データにおいては，ストップワードと記号文字のように性質の異なる無関連オブジェクトの存在が考えられる．また，SNS におけるユーザネット



(a) 入力　　(b) 出力

図 1: 提案法の概要図

ワークの関係データについても，ユーザをランダムにフォローするような Bot ユーザや活動を停止しているため誰もフォローしないユーザのように，異なる性質を持つ無関連オブジェクトの存在が考えられる．このように，実世界データには同じ関係データ内に性質の異なる無関連オブジェクトが含まれることが考えられる．

そこで本研究では，関係データに多様な無関連オブジェクトが含まれる可能性を考慮した新たな確率モデルを提案する．図 1 に提案法によるクラスタリングの例を示す．提案法は，関係データを入力として，関係データを構成するオブジェクト集合のクラスタリング結果とクラスタ毎の関係確率を出力する．左端の行列は提案法の入力となる関係データを表しており，行列の各要素は関係データを構成するデータエントリに対応する．i 行 j 列が黒いときは，オブジェクト i とオブジェクト j 間に関係が存在することを表し，白いときは両者間に関係がないことを表す．ここで，行方向のインデックスと列方向のインデックスは異なるオブジェクトを指すことに注意する．中央の行列は，提案法によって推定された各クラスタ間の関係確率であり，A, B, C, D と a, b, c はそれぞれ行方向と列方向のオブジェクトのクラスタのインデックスを表している．行列中の各要素の色の濃淡はクラスタ間の関係確率に対応しており，より黒い色はより高い関係確率を表す．このようなクラスタ間の関係確率を推定することにより，提案法は未観測な関係データのデータエントリについてもその関係の有無を予測することができる．ここで，クラスタ A, B は相手クラスタによらず関係確率が一定なクラスタであることに注意する．提案法では無関連オブジェクトをこのようなクラスタへと割り当てることで，無関連オブジェクトをクラスタリングの対象から除外する．また性質の異なる無関連オブジェクトが複数あった場合でも，それらを割り当てるために必要な数だけこのようなクラスタを生成できる点が提案法の特徴である．右端の行列は，提案法が出力するクラスタリングの結果から，同じクラスタのオブジェクトが隣り合うように行と列を並び替えた関係データであり，どのオブジェクトがどのクラスタへと割り当てられたかを表している．例えば，行方向のオブジェクト 3 と 5 はクラスタ C に，列方向のオブジェクト 4 と 6 はクラスタ b に割り当てられていることが読み取れる．またクラスタの割り当て結果から，各オブジェクトが関連オブジェクトか否かを読み取ることもでき，例えば行方向のオブジェクト 1 と 6 は，クラスタ A, B に割り当てられていることから，提案法により無関連オブジェクトだと推定されたことがわかる．関係データを区切っている赤い線はクラスタ間の境界であり，赤い線で区切られた各ブロックは同じ関係確率から生成されたと推定されたことを表している．ここで，クラスタ間の境界線による関係データの区切りは，推定されたデータの潜在構造に対応する．

以下の本文では，まず 2 章で関連研究を説明し，3 章で本稿で扱う関係データと提案モデルの基盤となる IRM について説明する．4 章では提案法における確率モデルと推論について説明し，5 章では人工データと実データによる実験とその結果について述べる．最後に 6 章で結論と今後の展望を述べる．

2. 関 連 研 究

確率的ブロックモデルの枠組みを用いた確率モデルはさかんに研究されており，様々な手法が提案されている [4] [7] [10] [11] [12] [11] [13]．これらの中で，特に提案法と類似する手法としては，オブジェクトを 1 つのクラスタへと割り当てるモデル [4] [11] とオブジェクトを関連オブジェクトと無関連オブジェクトに区分するモデル [7] [10] が挙げられる．IRM [4] や IRM を拡張した Bayesian Community Detection [11] は，オブジェクトをそれぞれの 1 つのクラスタへと割り当てる点で提案法と類似する．しかしこれらの手法は関連オブジェクトと無関連オブジェクトを区別せずクラスタリングを行う．一方で提案法は，関連オブジェクトと無関連オブジェクトを区別し，それぞれを異なる種類のクラスタへと割り当てるため，これらの手

法とは位置づけが異なる．また SIRM [7] や Relevance-based Overlapping Clustering [10] は，オブジェクトを関連オブジェクトと無関連オブジェクトに区分した上でクラスタリングを行うという点で提案法と類似する．しかしこれらの手法では，全ての無関連オブジェクトに同一の関係確率を仮定するため，異なる性質を持つ無関連オブジェクトには対応することができない．一方で提案法は，無関連オブジェクトをその性質に応じて異なるクラスタへと割り当てることで多様な無関連オブジェクトに対応することができるため，これらの手法とは位置づけが異なる．

3. 準　　備

3.1 関係データ

まずは本稿で扱う関係データについて説明する．一般に関係データは行列やテンソルとして表現される．今，I 個のオブジェクトからなるオブジェクト集合 $\mathbf{D}_1 = \{1, 2, ..., I\}$ と J 個のオブジェクトからなるオブジェクト集合 $\mathbf{D}_2 = \{1, 2, ..., J\}$ について表現した $I \times J$ 行列 $\mathbf{X}$ を考える．ここで，$\mathbf{X}$ は要素 $x_{i,j} \in \{0, 1\}$ はオブジェクト i とオブジェクト j の関係の有無を表現し，値が 1 のときは関係が存在することを，0 のときは関係が存在しないことを示す．またこのとき，$\mathbf{X}$ をオブジェクト集合 $\mathbf{D}_1, \mathbf{D}_2$ についての関係データと呼ぶ．購買履歴データを例にとると，$\mathbf{D}_1$ には客の集合，$\mathbf{D}_2$ には商品の集合，$\mathbf{X}$ の各要素 $x_{i,j}$ には客 i が商品 j を購入したかどうかを 0, 1 で表した記録をそれぞれ対応させることができる．本稿では 2 つのドメイン間の関係（$\mathbf{D}_1 \times \mathbf{D}_2$）を主に扱うが，単一ドメイン間の関係（$\mathbf{D}_1 \times \mathbf{D}_1$）や，3 ドメイン間の関係（$\mathbf{D}_1 \times \mathbf{D}_2 \times \mathbf{D}_3$）等，任意の関係データを提案法は扱うことができる．

3.2 無限関係モデル

本節では提案法のベースモデルとなる無限関係モデル（IRM）について説明する．IRM は，関係データの生成過程をモデリングした確率的生成モデルである SBM の拡張手法であり，ディリクレ過程を用いることで分割するクラスタ数を自動で推定できる点に特徴がある．IRM は関係データ $\mathbf{X}$ の各要素 $x_{i,j}$ を以下のように生成する．

$$
\begin{aligned}
x_{i,j} | Z, \Theta \sim \text{Bernoulli}(\theta_{z_{1,i}\, z_{2,j}}) \qquad (1)\\
i = 1, ..., I, \quad j = 1, ..., J, \quad 0 \leqq \theta_{k,\ell} \leqq 1\\
z_{1,i} \in \{1, ..., \infty\}, \quad z_{2,j} \in \{1, ..., \infty\}
\end{aligned}
$$

ここで，$\theta_{k,\ell}$ はクラスタ k, ℓ 間の関係確率を，$z_{1,i}$ はドメイン 1 のオブジェクト i が所属するクラスタを，$z_{2,j}$ はドメイン 2 のオブジェクト j が所属するクラスタをそれぞれ表す変数である．式 1 はオブジェクト i, j 間の関係 $x_{i,j}$ がそれぞれの所属クラスタ $z_{1,i}, z_{2,j}$ 間の関係確率 $\theta_{z_{1,i}\, z_{2,j}}$ をパラメータとする Bernoulli 分布から生成されることを示している．

4. 提　案　法

4.1 提案モデル

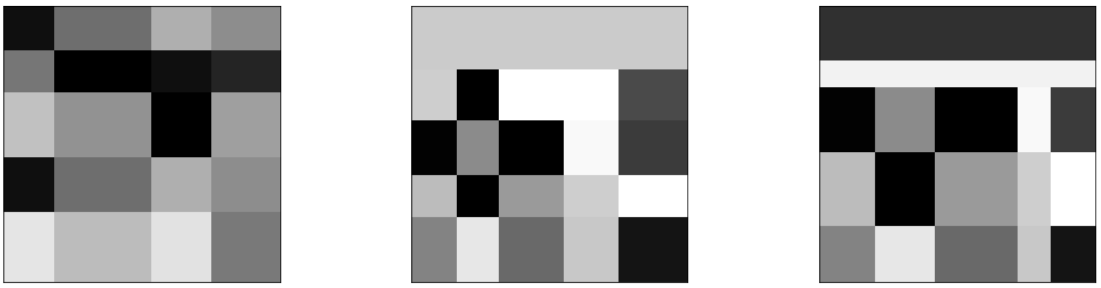
提案法は IRM 同様に確率的生成モデルとして表現される．提案モデルでは，オブジェクトの所属するクラスタとして，相手クラスタによって関係確率の変わる従来のクラスタに加えて，相手クラスタによらず関係確率が一定なクラスタを導入する．本稿では前者を関連クラスタ，後者を無関連クラスタと呼ぶ．無関連クラスタには無関連オブジェクトを割り当て，関連クラスタには関連オブジェクトを割り当てることで，オブジェクト集合から関連オブジェクトと無関連オブジェクトを分離し，関連オブジェクトについてのクラスタリングを頑健に行う点に提案法の特徴がある．また，無関連オブジェクトを割り当てるためのクラスタを必要に応じて複数個生成することにより多様な無関連オブジェクトに対応できる点も提案法の特徴である．以下の説明では，簡単のためドメイン 1 にのみ無関連オブジェクトを認める場合を扱う．

提案モデルでは，各オブジェクトが関連性を表す潜在変数 $r_{1,i}$ を持つと仮定する．$r_{1,i}$ はドメイン 1 のオブジェクト i についての変数であり，$r_{1,i} = 1$ のときはオブジェクト i が関連オブジェクトであることを，$r_{1,i} = 0$ のときはオブジェクト i が無関連オブジェクトであることを表す．$r_{1,i} = 1$ のときは IRM 同様，オブジェクト i，　j が所属する関連クラスタ $z_{1,i}, z_{2,j}$ 間の関係確率 $\theta_{z_{1,i}\, z_{2,j}}$ をパラメータとする Bernoulli 分布から関係 $x_{i,j}$ を生成する．一方で，$r_{1,i} = 0$ のときはオブジェクト j の所属クラスタに依存しない関係確率 $\phi_{y_{1,i}}$ をパラメータとする Bernoulli 分布から関係 $x_{i,j}$ を生成する．ここで $y_{1,i}$ はオブジェクト i の所属する無関連クラスタを表し，ϕ_m は無関連クラスタ m に所属するオブジェクトが他のオブジェクトと関係を持つ確率を表す．提案モデルではこのように，関連クラスタには相手オブジェクトの所属クラスタごとに異なる関係確率 $\theta_{k,\ell}$ を持たせるのに対して，無関連クラスタには相手オブジェクトの所属クラスタに依存しない一定の関係確率 ϕ_m を持たせることで．無関連オブジェクトが含まれる関係データの生成を表現する．

以下に提案モデルの生成モデル全体を示す．

$$
\begin{aligned}
&\theta_{k,\ell} \sim \text{Beta}(a, b) && 0 \leqq \theta_{k,\ell} \leqq 1,\ k, \ell \in \{1, ..., \infty\}\\
&\phi_m \sim \text{Beta}(c, d) && 0 \leqq \phi_m \leqq 1,\ m \in \{1, ..., \infty\}\\
&\lambda_1 \sim \text{Beta}(e, f) && 0 \leqq \lambda_1 \leqq 1\\
&r_{1,i} \sim \text{Bernoulli}(\lambda_1) && i = 1, ..., I,\ r_{1,i} \in \{0, 1\}\\
&z_{1,i} | r_{1,i} = 1 \sim \text{CRP}(\alpha_1) && i = 1, ..., I,\ z_{1,i} \in \{1, ..., \infty\}\\
&y_{1,i} | r_{1,i} = 0 \sim \text{CRP}(\beta) && i = 1, ..., I,\ y_{1,i} \in \{1, ..., \infty\}\\
&z_{2,j} \sim \text{CRP}(\alpha_2) && j = 1, ..., j,\ z_{2,j} \in \{1, ..., \infty\}\\
&x_{i,j} | Z, R, \Theta, \Phi \sim \text{Bernoulli}(\theta^{r_{1,i}}_{z_{1,i}, z_{2,j}} \phi^{1-r_{1,i}}_{y_{1,i}})
\end{aligned}
$$

ここで，$a, b, c, d, e, f, \alpha_1, \alpha_2, \beta$ はハイパーパラメータ，λ_1 はドメイン 1 のオブジェクトが関連オブジェクトである確率を表す変数である．また CRP はノンパラメトリックベイズの枠組みを用いた確率過程である Chinese Restaurant Process [8] で，任意の個数のオブジェクト集合を分割する際に考えられる各分割について確率を与える確率分布である．この確率過程を用いることで提案法ではクラスタ数をパラメータとして与える必要



(a) IRM　　(b) SIRM　　(c) 提案法

図 2: 各モデルで仮定される関係データの潜在構造

なく，関連オブジェクト集合に加えて無関連オブジェクト集合についても，分割するクラスタ数を自動で推定することができる．ここで，CRP からは正の整数が離散値として生成されるものとし，$y_{1,i}=0$ if $r_{1,i}=1$，$z_{1,i}=0$ if $r_{1,i}=0$ とする．なお上述の生成モデルにおいてはドメイン 1 にしか無関連オブジェクトの存在を認めなかったが，仮に 2 つ以上のドメインに無関連オブジェクトの存在を認めそれぞれのドメインに異なる無関連クラスタを認める場合には，無関連クラスタ間の関係確率について，どれかの関係確率の 1 つを採用するといった，新たな定義を与える必要があることに注意する．

4.1.1 既存手法との関係

提案モデルは IRM に無関連オブジェクトと無関連クラスタの生成過程を追加したモデルである．無関連オブジェクトと無関連クラスタの存在を仮定しない場合，すなわち全てのオブジェクトが関連オブジェクトであると仮定するとき，提案モデルは IRM と同等のモデルとなる．具体的には上述の生成モデルにおいて $\lambda_1=1$ の場合がそれにあたる．

また，オブジェクト集合を関連オブジェクトと無関連オブジェクトに分割し，関連オブジェクト集合においてクラスタ分割を行う SIRM についても，無関連クラスタのクラスタ数を 1 に限定した場合の提案モデルとみなすことができる．具体的には上述の生成モデルにおいて $y_{1,i}=1$ if $r_{1,i}=0$ の場合がそれにあたる．ただし SIRM は無関連オブジェクトを割り当てるためのクラスタを 1 種類に限定し複数のドメイン間で共有する．

図 2 に IRM と SIRM と提案モデルにおいて仮定される関係データの潜在構造の模式図を示す．IRM では全てのクラスタの組についてそれぞれ関係確率が別々に定義されるようなブロック構造しか仮定されないのに対して，SIRM や提案モデルでは無関連クラスタを導入することで，相手のクラスタによって関係確率が変わらないようなクラスタの存在を認める潜在構造を仮定できる．また SIRM が潜在構造中に無関連クラスタを 1 種類しか仮定できないのに対して，提案モデルでは任意の個数の無関連クラスタを持つ潜在構造を仮定できるという違いがある．

4.2 変数推論

本節では，提案モデルによってクラスタリングを行うアルゴリズムについて説明する．提案モデルでは，入力となる関係データ $\mathbf{X}$ を観測値として，未観測の変数 Y,Z,R を Gibbs サンプリングを用いて推論することによって各オブジェクトのクラスタ割り当てを推定する．変数 Φ,Θ,λ は未観測であるが，これらは積分消去が可能であるためサンプリングの対象とはしない．また Gibbs サンプリングは変数についての事後分布の近似をサンプリングによる経験分布を用いて求める手法であるが，本稿ではサンプリングを適当な回数繰り返した時点での変数の値を推論値としてそのまま用いる．

以下の節では，それぞれのドメインにおいてサンプリングに用いる条件付き確率について説明する．なお提案モデルは無関連オブジェクトの存在を認めるドメイン 1 と無関連オブジェクトの存在を認めないドメイン 2 で非対称なため，条件付き確率の式も異なることに注意する．

4.2.1 ドメイン 1 における変数のサンプリング

ドメイン 1 には無関連オブジェクトの存在を認めるため，オブジェクト i について求めたい変数は，$r_{1,i}$ と所属クラスタを表す $y_{1,i}$，$z_{1,i}$ である．本節ではこれらの変数を同時にサンプリングすることを考え，これらの変数の同時確率の条件付き確率について説明する．オブジェクト i についての変数の条件付き確率は以下のように書ける．

$$
\begin{aligned}
&p\left(z_{1,i}=k,y_{1,i}=m,r_{1,i}=n \mid \mathbf{X},Z^{\backslash i},Y^{\backslash i},R^{\backslash i}\right)\\
\propto&\frac{p(\mathbf{X}|\ z_{1,i}=k,y_{1,i}=m,r_{1,i},Z^{\backslash i},Y^{\backslash i},R^{\backslash i})}{p(\mathbf{X}^{\backslash i}|\ Z^{\backslash i},Y^{\backslash i},R^{\backslash i})}\\
\times&\frac{p(z_{1,i}=k,y_{1,i}=m,Z^{\backslash i},Y^{\backslash i}|\ r_{1,i}=n,R^{\backslash i})}{p(Z^{\backslash i},Y^{\backslash i}|\ R^{\backslash i})}\frac{p(r_{1,i}=n,R^{\backslash i})}{p(R^{\backslash i})}
\end{aligned}
$$

右辺の各項はそれぞれ以下のように表せ，これらは容易に計算できる．

$$
\frac{p(r_{1,i}=n,R^{\backslash i})}{p(R^{\backslash i})}=\begin{cases}e+\sum_{i'\neq i}r_{1,i'} & (n=1)\\ f+\sum_{i'\neq i}(1-r_{1,i'}) & (n=0)\end{cases}
$$

$$
\begin{aligned}
&\frac{p(z_{1,i}=k,y_{1,i}=m,Z^{\backslash i},Y^{\backslash i}|\ r_{1,i}=n,R^{\backslash i})}{p(Z^{\backslash i},Y^{\backslash i}|\ R^{\backslash i})}\\
=&\begin{cases}
\dfrac{\beta}{\beta+\sum_{m'}\sum_{i'\neq i}(1-r_{i'})\mathbb{I}(y_{1,i'}=m')} & (m=M+1,k=0)\\
\dfrac{\sum_{i'\neq i}(1-r_{i'})\mathbb{I}(y_{1,i'}=m)}{\beta+\sum_{m'}\sum_{i'\neq i}(1-r_{i'})\mathbb{I}(y_{1,i'}=m')} & (1\leqq m\leqq M,k=0)\\
\dfrac{\sum_{i'\neq i}r_{i'}\mathbb{I}(z_{1,i'}=k)}{\alpha_1+\sum_{k'}\sum_{i'\neq i}r_{i'}\mathbb{I}(z_{1,i'}=k')} & (1\leqq k\leqq K,m=0)\\
\dfrac{\alpha_1}{\alpha_1+\sum_{k'}\sum_{i'\neq i}r_{i'}\mathbb{I}(z_{1,i'}=k')} & (k=K+1,m=0)
\end{cases}
\end{aligned}
$$

$$\frac{p(\mathbf{X}|\ z_{1,i}=k, y_{1,i}=m, r_{1,i}=n, Z^{\setminus i}, R^{\setminus i})}{p(\mathbf{X}^{\setminus i}|\ Z^{\setminus i}, R^{\setminus i})} = \begin{cases} \dfrac{B(c+U_m^+, d+U_m^-)}{B(c+U_m^{+\setminus i}, d+U_m^{-\setminus i})} & (r_{1,i}=0) \\ \displaystyle\prod_{\ell'} \dfrac{B(a+T_{k,\ell'}^+, b+T_{k,\ell'}^-)}{B(a+T_{k,\ell'}^{+\setminus i}, b+T_{k,\ell'}^{-\setminus i})} & (r_{1,i}=1) \end{cases}$$

where

$$T_{k,\ell}^+ = \sum_i \sum_j r_{1,i} x_{i,j} \mathbb{I}(z_{1,i}=k)\mathbb{I}(z_{2,j}=\ell)$$

$$T_{k,\ell}^- = \sum_i \sum_j r_{1,i} (1-x_{i,j}) \mathbb{I}(z_{1,i}=k)\mathbb{I}(z_{2,j}=\ell)$$

$$U_m^+ = \sum_i \sum_j (1-r_{1,i}) x_{i,j} \mathbb{I}(y_{1,i}=m)$$

$$U_m^- = \sum_i \sum_j (1-r_{1,i})(1-x_{i,j}) \mathbb{I}(y_{1,i}=m)$$

ここで，M, K はそれぞれ各時点での関連クラスタ，無関連クラスタの数である．k, m, n の値に応じてこれらの値を求めることで，変数の条件付き確率を求めることができる．各変数について取り得る値のそれぞれについて条件付き確率を求めることでその変数についての確率分布が求まり，求めた分布から変数のサンプリングが可能となる．これを全てのオブジェクトの各変数について繰り返すことで変数のサンプリング列が求まり，この経験分布を以って事後分布は近似される．

4.3 ドメイン 2 における変数のサンプリング

ドメイン 2 のオブジェクトについては求めたい変数が所属クラスタを表す Z だけであるため，オブジェクト j についての変数 $z_{2,j}$ の条件付き確率は以下のようになる．

$$p\ (z_{2,j}=\ell \mid \mathbf{X}, Z^{\setminus j}, Y, R) = \frac{p\ (z_{2,j}=\ell, \mathbf{X}, Z^{\setminus j}, Y, R)}{p\ (\mathbf{X}, Z^{\setminus j}, Y, R)}$$

$$\propto \frac{p(\mathbf{X}|\ z_{2,j}=\ell, Z^{\setminus j})}{p(\mathbf{X}^{\setminus j}|\ Z^{\setminus j})} \frac{p(z_j=\ell, Z^{\setminus j}|)}{p(Z^{\setminus j}|)}$$

右辺の各項は以下に示すように表せ，これらもまた計算可能である．後の手順はドメイン 1 におけるものと同様であるため割愛する．

$$\frac{p(z_{2,j}=\ell, Z^{\setminus j})}{p(Z^{\setminus j})} = \begin{cases} \dfrac{\sum_{j' \neq j} \mathbb{I}(z_{2,j'}=\ell)}{\alpha_2 + \sum_{\ell'} \sum_{j' \neq j} \mathbb{I}(z_{2,j'}=\ell')} & (1 \leqq \ell \leqq L) \\ \dfrac{\alpha_2}{\alpha_2 + \sum_{\ell'} \sum_{j' \neq j} \mathbb{I}(z_{2,j'}=\ell')} & (\ell = L+1) \end{cases}$$

$$\frac{p(\mathbf{X}|\ z_{2,j}=\ell, Z^{\setminus j})}{p(\mathbf{X}^{\setminus j}|\ Z^{\setminus j})} = \prod_{k'} \frac{B(a+T_{k',\ell}^+, b+T_{k',\ell}^-)}{B(a+T_{k',\ell}^{+\setminus j}, b+T_{k',\ell}^{-\setminus})}$$

5. 実　　験

5.1 データセット

実験では人工データと実世界データの両方を用いて提案モデルの有効性を検証する．人工データとしては，ドメイン 1 に関連オブジェクトを 80，無関連オブジェクトを 720，ドメイン 2 に関連オブジェクト 1000 を持つ 2 ドメイン間の関係データを用いる．人工データはドメイン 1 に関連クラスタが 4，無関連クラスタが 12，ドメイン 2 に関連クラスタが 50 あるような潜在構造を持つように生成する．実世界のデータとしては，20news(注1)のデータセットと Google+のデータセット [9] から構築した 2 ドメイン間の関係データを用いる．20news のデータセットは，20 のニュースグループの文書データをおよそ同数ずつ計約 2 万件収集したデータセットである．実験では，このデータセットからランダムに 1000 文書を抽出し，それらの文書中で出現頻度が上位 1000 の単語を抽出し，どの文書にどの単語が出現したかを表現する関係データを作成する．Google+のデータセットは Google+上の 132 ユーザについて，そのユーザがフォローする全てのユーザについて互いのフォロー関係の有無を取得したユーザネットワークデータのセットである．実験ではこのデータセットからユーザネットワークを 1 つ選択し，ネットワークに含まれるユーザについてフォローする側とフォローされる側という 2 側面をそれぞれ別々のドメインのオブジェクト集合として誰が誰をフォローしているかを表現する関係データを作成する．また Google+の実験データについては，フォローする側のオブジェクト集合に，相手ユーザをその所属クラスタによらず一定の割合でフォローするユーザという，仮想的な無関連オブジェクトを追加したデータも作成する．追加する無関連オブジェクトはそれぞれ 0.15，0.30，0.45，0.60 という関係確率を持つ 4 種類の無関連クラスタから 5 つずつ生成する．

5.2 評 価 方 法

本稿ではリンク予測のタスクを通してモデルの有効性を評価する．具体的には，各実験データについて全体の 20 %のデータエントリをテストデータとして隠した状態で，残りのデータを学習データとしてクラスタリングを行い，クラスタ割り当ての結果から各データエントリにおける関係の有無をどの程度推定できるかを評価する．評価指標には ROC(Receiver Operating Characteristic) 曲線の AUC（Area Under Curve）を用いる．またテストデータに対するモデルの尤度も評価に用いる．比較手法には提案法の基盤となる IRM と無関連オブジェクトを考慮した既存手法である SIRM を用いる．各クラスタ間の関係確率については，クラスタリングの結果得られた変数の値を用いて，その事後確率を最大化する値として求める．

また Google+の実験データに人為的に無関連オブジェクトを加えたデータについては，上記の指標に加えて，追加した無関連オブジェクトのうち無関連クラスタへと割り当てることができたオブジェクトの数を評価する．この実験の目的は，提案

(注1)：http://qwone.com/~jason/20Newsgroups/

表 1: 20news における単語の無関連クラスタ例

クラスタ番号	単語
1	newsgroups: path: from: subject: message-id:
2	state (usenet thu, tue, tell mon, & 18
3	university, system) ... / institute wanted center data * article-i.d.: sat, price mail rec.autos space canada

表 2: AUC

	IRM	SIRM	提案法
synth	0.9215	0.9218	**0.9242**†
20news	0.8594	0.8594	**0.8603**†
Google+	0.9563	0.9563	**0.9566**
Google+(w/noise)	0.9510	0.9515	**0.9520**†

法による無関連オブジェクトの検出性能を評価することと，提案法が効果的な無関連オブジェクの性質を明らかにすることである．

実験では各データにつき学習データとテストデータの組み合わせを 10 通り作成し，それぞれの組み合わせについて各モデルにつき 10 試行クラスタリングを行い，それぞれのモデルにおいて最も学習データに対する尤度が高くなった試行の結果をその組み合わせについてのリンク予測に用いる．モデル間の比較においては，各データにつき 10 通りのリンク予測結果を用いる．また，各データにおける学習データとテストデータの組み合わせは適用する手法間で共通のため，有意差の検定には対応ありの t 検定を用いる．なお各手法におけるクラスタリングでは，変数推論において Gibbs サンプリングを各変数につき 500 回行う．

5.3 実験結果

まず，図 3，4 に人工データと 20news のデータに対する各手法によるクラスタリング結果を示す．図中のグラフの見方は図 1 におけるものに順ずる．これらの結果から提案モデルが，多様な無関連オブジェクトをそれぞれの関係確率に応じて異なる無関連クラスタへと割り当てることで，比較手法より精緻なクラスタリングが実現できていることを確認できる．

表 1 に，20news のデータに対して提案法を適用した結果得られた単語の無関連クラスタの例を示す．表中の左列はクラスタ番号を，右の列はそのクラスタに割り当てられた単語を示している．表より，クラスタ 1 にはネットニュースである 20news のどの文書にもヘッダとして出現するような単語が，クラスタ 2 には曜日を示すような単語などが，クラスタ 3 には記号文字などがそれぞれ抽出できていることが確認できる．ここで抽出されている単語はいずれも文書の種類と関係なく出現するようなものであることから，提案法が狙い通り無関連オブジェクトの多様性に対応できることが示唆されたと考える．

次に，表 2 にリンク予測の実験結果を，表 3 にテストデータに対するモデルの対数尤度を示す．ここでテストデータ対数尤度とは，それぞれのモデルにおいて各変数が推論の結果得られた値をとる場合の，テストデータに対するモデルの対数尤度である．表中で，各行は左端の列に示されるデータセットに対する各モデルでの実験結果に対する評価指標の平均値を示す．また，太字となっている値は各行で最も高い値を，その中でも † が付いている値は他の値よりも，検定により有意水準 5%で，有意に高いことが示された値を示している．結果より，どのデータセット，どの指標についても提案法が最も高い値を示し．Google ＋のデータ以外については，提案法が有意に高い値を示すことを確認した．このことより，今回作成した人工データや人為的に無関連オブジェクトを追加した Google ＋のデータのような，多様な無関連オブジェクトを含むデータにおいては提案法が有用であることが示されたと考えられる．また無関連オブジェクトを追加しなかった Google ＋のデータについても有意な差は出なかったものの，提案法が少なくとも比較手法と同等以上の性能を発揮することを確認できた．このことから，多様な無関連オブジェクトを含まないようなデータに対しても提案法が IRM や SIRM と同様に有効性を持つことが示されたと考えられる．

表 3: テストデータ対数尤度

	IRM	SIRM	提案法
synth	-33977	-33901	**-33387**†
20news	-35422	-35420	**-35370**†
Google+	-7332	-7296	**-7295**
Google+(w/noise)	-8581	-8546	**-8481**†

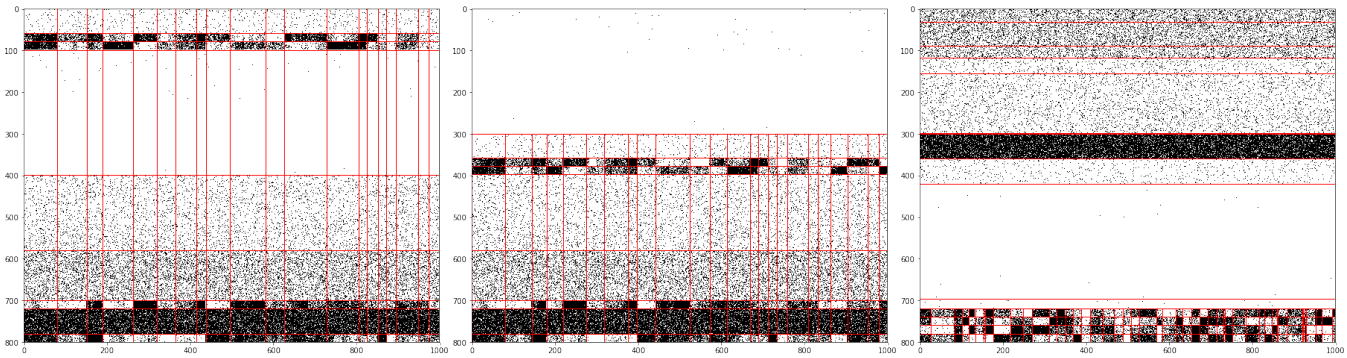
表 4: 無関連オブジェクトの抽出結果

	SIRM	提案法
人工 1（0.15）	1	5
人工 2（0.30）	0.5	5
人工 3（0.45）	1	5
人工 4（0.60）	0.5	5

また，表 4 に Google+のデータに無関連オブジェクトを追加したデータからの無関連オブジェクトの抽出結果を示す．なお IRM は無関連オブジェクトを明示的に抽出できないため，比較をしていない．表中の左端の列はそれぞれの手法で抽出できた無関連オブジェクトの種類を表しており，括弧内の数字はそれらのオブジェクトが持つ関係確率を表している．中央列と右端の列はそれぞれの手法において抽出できたオブジェクト数の平均値を表している．結果より，提案法が人為的に付与した全ての無関連オブジェクトについて，関係確率の大小によらず高い精度で抽出できていることが確認できる．これは SIRM ではほとんどの無関連オブジェクトを抽出できていないという結果と対照的であり，提案法が関係データからの無関連オブジェクト検出に有効であることを示していると考えられる．

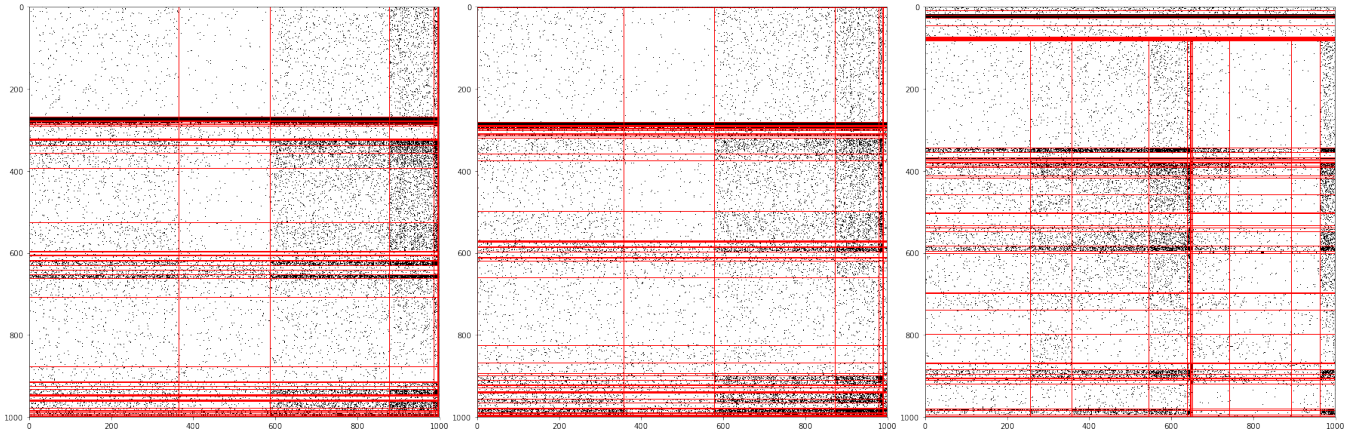
6. おわりに

本稿では，性質が異なる複数のノイズが含まれるような関係データに対しても，クラスタを精度よく抽出できる新たな確率モデルを提案した．提案法は，相手のクラスタに依存しない関



[1] IRM　　[2] SIRM　　[3] 提案モデル

図 3: 人工データのクラスタ割り当て結果



[1] IRM　　[2] SIRM　　[3] 提案モデル

図 4: 20news データのクラスタ割り当て結果

係確率を持つようなオブジェクトを，無関連オブジェクトとして性質に応じて異なる無関連クラスタへと割り当てることで，多様なノイズをクラスタリングの対象から分離できるのが特徴である．人工データを用いた実験では提案モデルが比較モデルに対して有意な性能の向上を示し，多様なノイズを含むデータにおける提案モデルの有効性が確認された．また実世界の関係データを用いた実験においても提案モデルは比較手法に対して同等以上の性能を示し，提案モデルの汎用的な有効性が確認された．今後の展望としては，既に SIRM の有効性が確認されているタスク [14] への提案モデルの適用や，時系列データに対応する IRM [15] 同様に時系列データを扱えるよう提案モデルを拡張することが考えられる．

文　　献

[1] K. Nowicki and T.A.B. Snijders, “Estimation and prediction for stochastic blockstructures,” Journal of the American Statistical Association, vol.96, no.455, pp.1077–1087, 2001.

[2] A. Goldenberg, A.X. Zheng, S.E. Fienberg, E.M. Airoldi, et al., “A survey of statistical network models,” Foundations and Trends® in Machine Learning, vol.2, no.2, pp.129–233, 2010.

[3] M.N. Schmidt and M. Morup, “Nonparametric bayesian modeling of complex networks: An introduction,” IEEE Signal Processing Magazine, vol.30, no.3, pp.110–128, 2013.

[4] C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda, “Learning systems of concepts with an infinite relational model,” AAAI, vol.3, p.5, 2006.

[5] B. Karrer and M.E. Newman, “Stochastic blockmodels and community structure in networks,” Physical Review E, vol.83, no.1, p.016107, 2011.

[6] T. Iwata, J.R. Lloyd, and Z. Ghahramani, “Unsupervised many-to-many object matching for relational data,” IEEE transactions on pattern analysis and machine intelligence, vol.38, no.3, pp.607–617, 2016.

[7] K. Ishiguro, N. Ueda, and H. Sawada, “Subset infinite relational models,” Artificial Intelligence and Statistics, pp.547–555, 2012.

[8] D. Blackwell and J.B. MacQueen, “Ferguson distributions via pólya urn schemes,” The annals of statistics, pp.353–355, 1973.

[9] J. McAuley and J. Leskovec, “Learning to discover social circles in ego networks,” Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, pp.539–547, NIPS’12, Curran Associates Inc., USA, 2012.

[10] J.J. Whang, P. Rai, and I.S. Dhillon, “Stochastic blockmodel with cluster overlap, relevance selection, and similarity-based smoothing,” Data Mining (ICDM), 2013 IEEE 13th International Conference onIEEE, pp.817–826 2013.

[11] M. Mørup and M.N. Schmidt, “Bayesian community detection,” Neural computation, vol.24, no.9, pp.2434–2456,

2012.
[12] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing, "Mixed membership stochastic blockmodels," Journal of Machine Learning Research, vol.9, no.Sep, pp.1981–2014, 2008.
[13] K. Miller, M.I. Jordan, and T.L. Griffiths, "Nonparametric latent feature models for link prediction," Advances in neural information processing systems, pp.1276–1284, 2009.
[14] T. Iwata and K. Ishiguro, "Robust unsupervised cluster matching for network data," Data Mining and Knowledge Discovery, vol.31, no.4, pp.1132–1154, 2017.
[15] K. Ishiguro, T. Iwata, N. Ueda, and J.B. Tenenbaum, "Dynamic infinite relational model for time-varying relational data analysis," Advances in Neural Information Processing Systems, pp.919–927, 2010.