修士論文

多様なノイズに頑健な関係データクラスタリング

武田 悠佑

2019年3月14日

奈良先端科学技術大学院大学
情報科学研究科

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

武田 悠佑

審査委員：

松本 裕治 教授	（主指導教員）
中村 哲 教授	（副指導教員）
澤田 宏 教授	（副指導教員）
岩田 具治 准教授	（副指導教員）
新保 仁 准教授	（副指導教員）

多様なノイズに頑健な関係データクラスタリング*

武田 悠佑

内容梗概

単語文書行列や SNS における友人関係ネットワークのようなノイズを多く含む関係データをクラスタリングするための確率モデルを提案する．多くの場合において関係データには，単語文書行列におけるストップワードや友人関係ネットワークにおけるスパムアカウントのような，他のオブジェクトとの間に有意な関係のパターンを示さないノイズオブジェクトが含まれている．従来の関係データクラスタリング手法では，全てのオブジェクトに対して有意な関係のパターンの存在を仮定するために，ノイズオブジェクトであってもいずれかのクラスタへと割り当ててしまい，それによってクラスタリング結果が悪影響を受けるという問題があった．提案法では，他のクラスタに対して特定のパターンを示さないノイズクラスタというクラスタを導入する．ノイズオブジェクトにはノイズクラスタを割り当てることで，他のオブジェクトからノイズオブジェクトを分離し，通常のオブジェクト間の関係のパターンをノイズを含む関係データから抽出できる．また提案法では多様なノイズを扱うために，無限個のノイズクラスタを仮定する．ディリクレ過程を用いることで，通常のクラスタとノイズクラスタの個数を与えられたデータから自動的に推定する．また提案モデルに対する推論アルゴリズムとして，崩壊型ギブスサンプリングに基づいた効率的なアルゴリズムを導出する．人工データと実世界データを用いた実験を通じて，提案法が従来法よりも高い精度で未観測データについて予測できることを示す．

キーワード

関係データ, クラスタリング, 教師なし学習, ノンパラメトリックベイズモデル

*奈良先端科学技術大学院大学 情報科学研究科 修士論文, 2019 年 3 月 14 日.

Robust Clustering for Relational Data with Various Noises*

Yusuke Takeda

Abstract

We propose a probabilistic model for clustering noisy relational data, such as document-word networks and friend links on social networking services. Relational data often contain noise objects, which do not have meaningful interaction patterns with other objects, such as stop words in document-word networks and spam accounts in friend networks. Existing clustering methods for relational data suffer from these noises because the noise objects are assigned into clusters by assuming that there exist meaningful interaction patterns. The proposed method contains noise clusters, which do not exhibit any patterns to other clusters, as well as normal clusters. The noise clusters enable us to separate the noise objects from the other objects, and help to extract patterns among the normal objects from noisy relational data. To handle various noises, the proposed method assumes an infinite number of noise clusters. Using Dirichlet processes, the number of noise and normal clusters are automatically estimated from the given data. We also present an efficient inference algorithm for the proposed model based on the collapsed Gibbs sampling. We demonstrate that the proposed model can more precisely predict the existence of unobserved relations than conventional models through the experiments using synthetic and real-world data sets.

Keywords:

Relational data, Clustering, Unsupervised learning, Nonparametric Bayesian model

*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, March 14, 2019.

目 次

1. はじめに **1**

2. 関連研究 **5**

2.1 関係データクラスリング 5

2.2 関連性推定手法 5

3. 準備 **8**

3.1 関係データ 8

3.2 確率的ブロックモデル 8

3.3 無限関係モデル 9

3.4 サブセット無限関係モデル 11

4. 提案手法 **13**

4.1 提案モデル 13

4.2 先行手法との関係性 15

4.3 推論 16

5. 実験 **22**

5.1 実験設定 22

5.2 実験結果 29

6. おわりに **34**

謝辞 **35**

参考文献 **36**

発表リスト **40**

図 目 次

1 提案法によるクラスタリング例 2
2 IRM のグラフィカルモデル 10
3 SIRM のグラフィカルモデル 12
4 提案モデルのグラフィカルモデル 15
5 先行手法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 17
6 人工データに対するクラスタリング結果 24
7 20News データに対するクラスタリング結果（行：単語，列：文書） 25
8 MovieLens データに対するクラスタリング結果（行：ユーザ，列：映画） 26
9 Google+データに対するクラスタリング結果（行：フォローする側，列：フォローされる側） 27
10 Google+(w/noise) データに対するクラスタリング結果（行：フォローする側，列：フォローされる側） 28

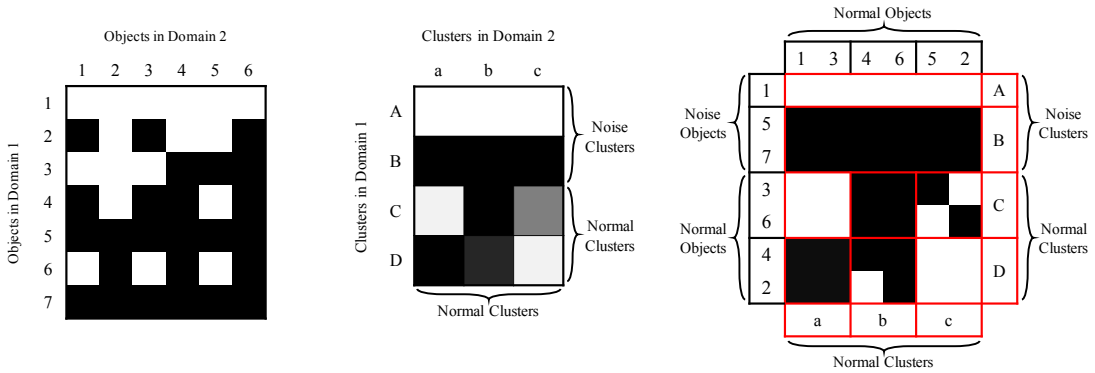
表 目 次

1 記号表記 16
2 リンク予測における評価データに対する AUC 23
3 評価データに対する 1 要素あたりの対数尤度 29
4 推定されたクラスタ割当と真のクラスタ割当の ARI 29
5 ノイズオブジェクトの検出率 30
6 人工データに対する性質別のノイズオブジェクト検出数 30
7 Google+(w/noise) データに対する性質別のノイズオブジェクト検出数 31
8 提案法により 20News データから検出されたノイズクラスタ例 . . 31
9 計算時間 (秒) 32

1. はじめに

SNS における友人関係ネットワークや，誰が何を買ったかという EC サイトでの購買履歴，どの文書にどの単語が出現したかという単語文書行列などのデータは，いずれもオブジェクト同士の関係性を表現する関係データである．関係データは今日社会のあらゆるところに存在し，それらをクラスタリングすること，すなわち関係データ内のオブジェクトを類似するオブジェクトのグループへと分割することは人工知能の重要な課題となっている．例えば，SNS 上の友人関係ネットワークにクラスタリングを適用すれば，ターゲットマーケティングやユーザ推薦に有用な類似したユーザのコミュニティを抽出できる．また，単語文書行列に対してクラスタリングを適用すれば，文書コーパスの分析に有用と考えられる文書のジャンルや単語のトピックが得られる．確率的ブロックモデル (stochastic block model; SBM) [25] は，関係データのクラスタリングのための代表的な確率的生成モデルであり．これまでに様々な領域での応用がなされてきた [6, 19, 29, 17]. SBM では，各オブジェクトにはいずれかの潜在クラスタが割り当てられており，各クラスタは他のクラスタそれぞれに対して特徴的な関係のパターンを有していると仮定する．この仮定に基づく SBM は，クラスタリングと同時にクラスタ間の関係のパターンを推定するため，未観測なデータについての予測にも応用できる [8]．例えば，購買履歴データに対して SBM を適用すれば，商品とユーザのグループを抽出すると同時にあるユーザがある商品を購入する確率を推定できる．

しかし，SBM は関係データに含まれるノイズを適切に扱えないという問題がある．ノイズとは他のオブジェクトとの間に有意な関係のパターンを示さないオブジェクトのことであり，本研究ではこれをノイズオブジェクトと呼ぶ．一方で，それ以外のオブジェクト，すなわち他のオブジェクトとの間に有意な関係のパターンを持つオブジェクトについては通常オブジェクトと呼ぶ．多くの場合において関係データはノイズオブジェクトを含んでいる．例えば，単語文書行列においては，ほとんどの文書に登場するストップワードや，文書の種類を問わずに出現する記号文字などはノイズオブジェクトといえる．SNS の友人関係ネットワークにおいてよく見つかる，ランダムにユーザをフォローするスパムアカウントなどもノイズオブジェクトとみなせる．SBM では，このようなノイズオブジェクトに対



(a) 関係データ　　(b) 関係確率　　(c) クラスタリング結果

図 1: 提案法によるクラスタリング例

しても有意な関係のパターンの存在を仮定していずれかのクラスタを割り当ててしまうため，ノイズオブジェクトを含む関係データからは潜在的な関係のパターンを適切に抽出できないという問題がある．

この問題を解決するために，サブセット無限関係モデル（subset infinite relational model; SIRM） [15] は，各オブジェクトが通常オブジェクトかノイズオブジェクトかを示す潜在変数を導入している．潜在変数の値を明示的に推論することで，SIRM はノイズオブジェクトと通常オブジェクトを分離し，頑健なクラスタリングを実現している．一方で，実世界の関係データには異なる性質を持つ様々な種類のノイズオブジェクトが含まれている場合がある．例えば，単語文書行列において，ストップワードと記号文字はともにノイズオブジェクトであるが，ストップワードは高い確率で出現すると考えられる一方，記号文字はストップワードとは異なるより低い確率で出現すると考えられる．また友人関係ネットワークにおいても，ランダムに他のユーザをフォローするスパムアカウントとほとんど誰もフォローしていない活動を停止したユーザはともにノイズオブジェクトであるが，その性質は明らかに異なる．SIRM は全てのノイズオブジェクトが同一の性質を持つと仮定するため，このように様々なノイズオブジェクトが含まれる関係データを適切に扱えない．

本論文では，多様なノイズオブジェクトが含まれる関係データをクラスタリングするための確率モデルを提案する．提案モデルは無限関係モデル（infinite relational model; IRM） [20] をベースとする．IRM はディリクレ過程をクラスタ割り当ての事前分布として用いることでクラスタの個数を自動的に推定する SBM の拡張手法であり，提案モデルでもクラスタの個数を自動的に推定する．提案モデルでは，従来のクラスタに加えて，ノイズクラスタという新たな種類のクラスタを導入する．従来の通常クラスタは他のクラスタに対してそれぞれ異なる関係のパターンを示すクラスタであったのに対して，ノイズクラスタは他のクラスタとの間に関係のパターン示さない，ノイズオブジェクトのためのクラスタである．提案法を含むブロックモデルでは，クラスタが持つ関係のパターンを関係確率と呼ばれるそのクラスタのオブジェクトが他のクラスタのオブジェクトと関係を持つ確率によって表現する．提案モデルでも従来の SBM と同様に通常クラスタは相手のクラスタに応じて異なる関係確率を持つ．一方，ノイズクラスタには相手のクラスタによらない一定の関係確率を持たせることで，提案モデルではノイズクラスタが他のクラスタとの間に関係のパターンを示さないことを表現する．また多様なノイズを扱うため，提案モデルでは無限個のノイズクラスタを仮定する．ディリクレ過程を用いることで通常クラスタと同様にノイズクラスタについてもその個数を与えられたデータから自動的に推定する．提案モデルでは，データから自動的にノイズクラスタの個数を決定した上で，ノイズオブジェクトにはその性質に応じたノイズクラスタを割り当てる．通常オブジェクトとノイズオブジェクトを違う種類のクラスタへ割り当てることでそれぞれ分離し，関係データの潜在的な関係のパターンを適切に抽出する．提案モデルの変数推論については，関係確率とクラスタが出現する確率を解析的に積分消去した上でクラスタ割当をサンプリングする，崩壊型ギブスサンプリングに基づいた効率的な推論アルゴリズムを導出する．図 1 に提案法によるクラスタリングの一例を示す．この例では 2 ドメイン間の関係データを扱う．ここで，ドメインとはユーザ，アイテム，文書，単語といったオブジェクトの種類を指し，例えば購買履歴データはユーザドメインのオブジェクトとアイテムドメインのオブジェクト間の関係データとみなせる．(a) は観測された関係データを行列の形で表現したものであり，行と列のインデッ

クスはそれぞれのドメインにおけるオブジェクトを指している．また，各要素はオブジェクトの間の関係の有無を黒（有）と白（無）で表現している．(b）はクラスタリングの結果推定されるクラスタ間の関係確率を表しており，行と列のインデックスはそれぞれのドメインにおけるオブジェクトのクラスタを表現している．各要素の色の濃さは関係確率の値の大小を表現しており，より暗い色はより高い関係確率を表す．ノイズクラスタ A と B に着目すると，それぞれ相手クラスタに依らない一定の関係確率を持つことがわかる．提案法では，ノイズオブジェクトにはこのようなノイズクラスタを割り当てることでノイズオブジェクトを通常オブジェクトから分離する．(c）は提案法によるクラスタリングの結果を表しており，クラスタ割り当ての結果に基づいて（a）の関係データの行と列を並び替えたものである．この例ではドメイン 1 のオブジェクト 1 にはノイズクラスタ A を，オブジェクト 5 と 7 にはノイズクラスタ B を割り当てられており，ノイズオブジェクトを通常オブジェクトから分離できていることがわかる．

以後の本論文の構成は以下の通りである．まず 2 章で関連研究について述べる．3 章では，本論文で扱う関係データの表記法についてと提案モデルのベースとなる SBM，IRM，そして提案モデルと関連性が高い SIRM について説明する．4 章では，多様なノイズを含む関係データクラスタリングのための確率モデルを提案し，崩壊型ギブスサンプリングに基づく推論アルゴリズムを導出する．5 章では提案モデルの有効性を人工データと実世界データを用いた実験によって示し，最後に 6 章で結論と今後の展望を述べる．

2. 関連研究

2.1 関係データクラスリング

提案法と同様な確率的ブロックモデルに基づく関係データクラスタリング手法としては様々なモデルが提案されている．1 つのオブジェクトに 1 つのクラスタを割り当てる手法としては，クラスタの個数を自動的に推定する無限関係モデル（IRM） [20] や，IRM をコミュニティ検出に適応するよう拡張したモデル [24], IRM に行列分解の手法を取り入れたモデル [30] 等が提案されている．また，1 つのオブジェクトに複数のクラスタを割り当てる手法としては，複数のクラスタを確率的に割り当てるソフトなクラスタリングモデル [1] や，複数のクラスタを決定的に割り当てるハードなクラスタリングモデル [23, 28] が提案されている．しかし，これらのモデルはいずれもノイズオブジェクトを扱う仕組みを持たないため提案モデルとは区別される．なお提案モデルは 1 つのオブジェクトを 1 つのクラスタに割り当てる IRM を基礎としたモデルであるが，提案モデルにおける無限個のノイズクラスタという機構は複数のクラスタを割り当てるモデルに対しても導入可能である．

一方で，確率モデルを用いない関係データクラスタリング手法も数多く提案されている．行列の固有値分解を用いるスペクトラルクラスリングに基づく手法としては，二部グラフで表現される関係データを扱える手法 [5] や，複数のドメイン間の関係についての関係データを扱える手法 [21] が提案されている．また，Hathaway らは，1 つのオブジェクトに対して複数のクラスタについての関係性を割り当てるソフトな関係データクラスリング手法として，Fuzzy c-means ベースの手法を提案している [10, 9]．しかし，これらの手法もノイズオブジェクトを扱う仕組みを持たないため，提案モデルとは区別される．

2.2 関連性推定手法

提案法では，関係データの各要素をノイズに関連するデータか推定し，ノイズデータとそうでないデータに異なる生成過程を仮定することで頑健なクラスタリ

ングを実現する．また，提案法は通常オブジェクトを関係データからクラスタ構造を推定するのに役立つ特徴量として，ノイズオブジェクトをそうでない特徴量として区別しているともみなせる．これらの点において，観測データが着目したい情報に関連するかを推定して生成過程を区別する手法や，性能を向上させるためにタスクに関連する特徴量を推定する手法と提案法は関連する．

ノイズに関連するデータとそうでないデータで生成過程を区別する関係データクラスタリング手法はこれまでに複数提案されている．Ishiguro らは，ノイズオブジェクトと他のオブジェクトとの関係はオブジェクトに依存しない同一の生成過程から生成されると仮定した上で，各オブジェクトについて通常オブジェクトかノイズオブジェクトかを推定することで頑健なクラスタリングを実現するサブセット無限関係モデル（SIRM）を提案している [15]．また，SIRM と同様の機構を，1 つのオブジェクトに複数のクラスタを決定的に割り当てるモデルに導入した手法も提案されている [31]．一方で，Ohama らは，オブジェクトではなく関係データの各要素についてノイズに関連するかどうか推定する手法を提案している [26]．これらの手法ではいずれもノイズに関連するデータの生成過程を 1 種類に限定しており，異なる性質を持つ複数のノイズオブジェクトには対応できない．提案法は，現状知られている中で，複数のノイズクラスタを仮定する最初の関係データクラスタリング手法である．

複数のデータ点をクラスタリングする一般的なクラスタリングについても，特徴量やデータを区別する手法は様々なものが提案されている．Hoff は，二値で表現される多次元データに対するディリクレ過程混合モデルについて，クラスタごとに関連する特徴量を選択する機構を導入し特徴量の部分集合を用いてクラスタリングするモデルを提案している [11]．また，Guan らは，同様のモデルを実数値のデータに対しても適用できるよう拡張したモデルを提案している [7]．一方で，Dave らは，Fuzzy c-means クラスタリングにノイズクラスタを導入し，外れ値のデータをそのクラスタへ割り当てることで頑健なクラスタリングを実現している [4]．

また，関係データに対する確率モデルの一種であるトピックモデルについても，データの生成過程を区別するモデルは複数提案されている．Chemudugunta らは，

単語文書行列について，トピックと文書とコーパスのそれぞれが単語の生成過程を持つと仮定して，観測された単語がどの過程から生成されたか推定するモデルを提案している [3]．また，Iwata らは，映画に対するタグデータのようなコンテンツデータに対するアノテーションデータについて，コンテンツのトピックに関連する生成過程とコンテンツに関連しない生成過程の二つを仮定し，観測されたデータがどちらの過程から生成されたか推定するモデルを提案している [18]．これらの手法においても，複数のノイズクラスタを仮定したものはなく，多様なノイズオブジェクトを陽にモデル化する提案法には新規性があると考えられる．

3. 準備

3.1 関係データ

本節では，本論文における関係データの表記法を説明する．ここでは，異なる 2 ドメインのオブジェクト集合 $\mathbf{D}_1$，$\mathbf{D}_2$ 間における二値で表現される関係 $\mathbf{D}_1 \times \mathbf{D}_2 \rightarrow \{0,1\}^{I \times J}$ を考える．I と J はそれぞれドメイン 1 とドメイン 2 のオブジェクトの総数である．この関係について観測される関係データを $\mathbf{X} = \{x_{i,j} \in \{0,1\}; i = 1,...,I,\ j = 1,...,J\}$ と表記する．各要素 $x_{i,j}$ はオブジェクト i とオブジェクト j 間に関係が，$x_{i,j} = 1$ のときには有ることを，$x_{i,j} = 0$ のときには無いことを表す．例えば，購買履歴データの場合，$\mathbf{D}_1$ はユーザのリストに，$\mathbf{D}_2$ はアイテムのリストに，$x_{i,j}$ の値はユーザ i がアイテム j を購入したかどうかに対応する．本論文では簡便のため，上述の 2 ドメイン間について二値で表現される関係を扱うが，1 ドメイン間のデータ ($\mathbf{D}_1 \times \mathbf{D}_1$) や，3 ドメイン間のデータ ($\mathbf{D}_1 \times \mathbf{D}_2 \times \mathbf{D}_3$) のようなより高次元なデータ，あるいは実数値を持つ関係データ ($\mathbf{D}_1 \times \mathbf{D}_2 \rightarrow \mathbb{R}^{I \times J}$) に対しても提案法の拡張は容易に行える．

3.2 確率的ブロックモデル

本節では，提案モデルや比較手法の基礎である確率的ブロックモデル（SBM）について説明する．SBM は，関係データについての確率的生成モデルであり，関係データに含まれる複数のドメインのオブジェクトを同時にクラスタリングできる．SBM では，各オブジェクトには潜在クラスタが割り当てられており，各オブジェクト間の関係はそれぞれの潜在クラスタ間に定義される何らかの確率分布から生成されると仮定する．SBM ではこの仮定に基づき各オブジェクトの潜在クラスタを推定することで，異なるドメインのオブジェクトを同時にクラスタリングすることを実現する．本論文では二値で表現される関係について扱うため，潜在クラスタ間に定義される確率分布を Bernoulli 分布に限定して以下の議論を進める．関係データ X の SBM における生成モデルは次のように表される．なお，以

下の式で $A \sim \mathrm{B}$ とは，変数 A が確率分布 B からサンプリングされることを表す．

$$
\begin{aligned}
\pi_1 &\sim \mathrm{Dirichlet}(\alpha_1) && \pi_1 = \{\pi_{1,1}, ..., \pi_{1,K}\},\ \pi_{1,k} \geq 0,\ \sum_{k=1}^{K} \pi_{1,k} = 1, \\
\pi_2 &\sim \mathrm{Dirichlet}(\alpha_2) && \pi_2 = \{\pi_{2,1}, ..., \pi_{2,L}\},\ \pi_{2,\ell} \geq 0,\ \sum_{\ell=1}^{L} \pi_{2,\ell} = 1, \\
z_{1,i} &\sim \mathrm{Categorical}(\pi_1) && i = 1, ..., I,\ z_{1,i} \in \{1, ..., K\}, \\
z_{2,j} &\sim \mathrm{Categorical}(\pi_2) && j = 1, ..., J,\ z_{2,j} \in \{1, ..., L\}, \\
\theta_{k,\ell} &\sim \mathrm{Beta}(a, b) && 0 \leq \theta_{k,\ell} \leq 1,\ k \in \{1, ..., K\}, \ell \in \{1, ..., L\}, \\
x_{i,j} &\sim \mathrm{Bernoulli}(\theta_{z_{1,i}, z_{2,j}}) && i = 1, ..., I,\ j = 1, ..., J,\ x_{i,j} \in \{0, 1\}.
\end{aligned}
$$

ここで，π_d はドメイン d における潜在クラスタの混合割合であり，K と L はそれぞれドメイン 1 とドメイン 2 の潜在クラスタの総数である．$\theta_{k,\ell}$ はクラスタ k と ℓ の間に定義される Bernoulli 分布のパラメータであり，クラスタ k のオブジェクトとクラスタ ℓ のオブジェクトの関係確率に対応している．i と j はそれぞれドメイン 1 とドメイン 2 のオブジェクトを指しており，$z_{d,i}$ はドメイン d のオブジェクト i に割り当てられる潜在クラスタを表現している．また，$\Theta = \{\{\theta_{k,\ell}\}_{\ell=1}^{L}\}_{k=1}^{K}$，$Z = \{\{z_{1,i}\}_{i=1}^{I}, \{z_{2,j}\}_{j=1}^{J}\}$ である．この生成モデルでは，まずハイパーパラメータ α_d をパラメータとする Dirichlet 分布からドメイン d におけるクラスタの混合割合 π_d が生成される．次に，生成された π_d をパラメータとする Categorical 分布からドメイン d におけるオブジェクト i のクラスタ割当 $z_{d,i}$ が生成される．またハイパーパラメータ a, b を持つ Beta 分布からクラスタ k, ℓ 間の関係確率 $\theta_{k,l}$ が生成され，関係 $x_{i,j}$ がオブジェクト i と j の割当クラスタ $z_{1,i}, z_{2,j}$ 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．

3.3 無限関係モデル

無限関係モデル (IRM) は，各ドメインにおける潜在クラスタ数に無限個の仮定をおいた SBM の拡張手法である．IRM では，ディリクレ過程をクラスタ割り当ての事前分布として導入することで，SBM においてはハイパーパラメータとして

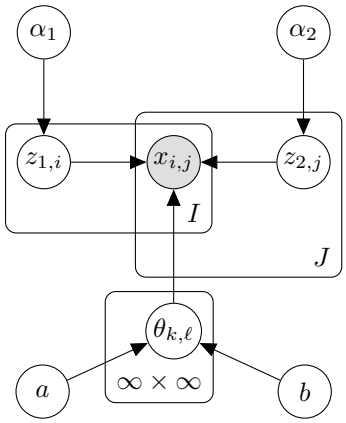


図 2: IRM のグラフィカルモデル

与える必要のあった K と L の値を与えられたデータから自動的に推定する．IRM における関係データ X の生成モデルは以下のように表される．

$$
\begin{aligned}
z_{1,i} &\sim \mathrm{CRP}(\alpha_1) && i = 1, ..., I,\ z_{1,i} \in \{1, ..., \infty\}, \\
z_{2,j} &\sim \mathrm{CRP}(\alpha_2) && j = 1, ..., J,\ z_{2,j} \in \{1, ..., \infty\}, \\
\theta_{k,\ell} &\sim \mathrm{Beta}(a, b) && 0 \leq \theta_{k,\ell} \leq 1,\ k, \ell \in \{1, ..., \infty\}, \\
x_{i,j} &\sim \mathrm{Bernoulli}(\theta_{z_{1,i}, z_{2,j}}) && i = 1, ..., I,\ j = 1, ..., J,\ x_{i,j} \in \{0, 1\}.
\end{aligned}
$$

SBM ではクラスタの混合割合 π_d を生成した上で，それを用いてクラスタ割当 $z_{d,i}$ を生成するのに対して，IRM ではクラスタ割当を Chinese Restaurant Process（CRP）[2] という確率過程から直接生成している．CRP とは，ディリクレ過程に基づくノンパラメトリックベイズの枠組みの確率過程の一種であり，オブジェクトの集合に対する可能な分割のそれぞれに対して確率を与える分布とみなせる．CRP を用いることで，暗黙的には無限個のクラスタからのクラスタ割当のサンプリングが可能となる．そのため，IRM では事前にクラスタ数を固定しなくても与えられたデータから適切なクラスタ数を自動的に推定できる．IRM のグラフィカルモデルを図 2 に示す．

3.4 サブセット無限関係モデル

サブセット無限関係モデル (SIRM) は，各オブジェクトが通常オブジェクトかノイズオブジェクトかを判定する機構を持った IRM の拡張手法である．SIRM では，ノイズオブジェクトと他のオブジェクトとの関係はそれぞれの潜在クラスタによらず同一の確率分布から生成されるという仮定に基づいて，関係データの生成過程を表現する．SIRM における関係データ X の生成モデルは次に示す式で表される．なお，SIRM はノイズオブジェクトの存在を特定のドメインに限定するようなモデルではないが，ここでは後述する提案モデルとの比較を容易にするため，ノイズオブジェクトはドメイン 1 にのみ存在すると仮定する場合の生成モデルを示す．

$$
\begin{aligned}
\lambda_1 &\sim \mathrm{Beta}(e,f) && 0 \leq \lambda_1 \leq 1, \\
r_{1,i} &\sim \mathrm{Bernoulli}(\lambda_1) && i = 1,...,I,\ r_{1,i} \in \{0,1\}, \\
z_{1,i} &\sim \mathrm{CRP}(\alpha_1) && i = 1,...,I,\ z_{1,i} \in \{1,...,\infty\}, \\
z_{2,j} &\sim \mathrm{CRP}(\alpha_2) && j = 1,...,J,\ z_{2,j} \in \{1,...,\infty\}, \\
\theta_{k,\ell} &\sim \mathrm{Beta}(a,b) && 0 \leq \theta_{k,\ell} \leq 1,\ k,\ell \in \{1,...,\infty\}, \\
\phi &\sim \mathrm{Beta}(c,d) && 0 \leq \phi \leq 1, \\
x_{i,j} &\sim \mathrm{Bernoulli}(\theta^{r_{1,i}}_{z_{1,i},z_{2,j}} \phi^{1-r_{1,i}}) && i = 1,...,I,\ j = 1,...,J,\ x_{i,j} \in \{0,1\}.
\end{aligned}
$$

IRM と比較すると SIRM では変数 $\lambda_1, r_{1,i}, \phi$ が新たに導入されている．ここで，変数 λ_1 はドメイン 1 のオブジェクト集合に含まれるノイズオブジェクトの割合，変数 ϕ はノイズオブジェクトと他のオブジェクトとの関係確率である．また，変数 $r_{1,i}$ はドメイン 1 の各オブジェクトがノイズオブジェクトかどうかを表し，オブジェクト i は，$r_{1,i} = 1$ のとき通常オブジェクト，$r_{1,i} = 0$ のときノイズオブジェクトとなる．SIRM では，オブジェクト i が通常オブジェクトかノイズオブジェクトかにより関係 $x_{i,j}$ の生成過程を分離する．具体的には，$r_{1,i} = 1$ の場合は SBM や IRM と同様にそれぞれの潜在クラスタ間に定義される関係確率をパラメータとする Bernoulli 分布から，$r_{1,i} = 0$ の場合は関係確率 ϕ をパラメータとする Bernoulli 分布から，それぞれ関係 $x_{i,j}$ が生成される．SIRM では，この生成モデルに基づいて各変数の値を推論することで，ノイズオブジェクトを含む関係データからノ

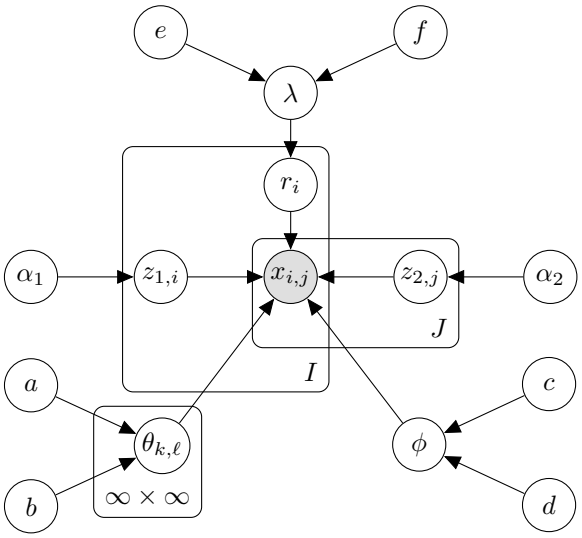


図 3: SIRM のグラフィカルモデル

イズオブジェクトを分離すると同時に通常オブジェクトをクラスタリングする.
SIRM のグラフィカルモデルを図 3 に示す.

4. 提案手法

4.1 提案モデル

提案モデルは，IRM を基礎として，ノイズクラスタという新たな種類のクラスタを導入した確率的生成モデルである．ノイズクラスタは，従来の通常クラスタとは異なり，他のクラスタとの間に相手のクラスタに依らない一定の関係確率を持つ，他のクラスタに対して関係のパターンを示さないクラスタである．また多様な種類のノイズオブジェクトを扱うため，ノイズクラスタの個数についても IRM における通常クラスタと同様に無限個の仮定を置く．提案法は，ノイズオブジェクトにはノイズクラスタを割り当て，通常オブジェクトには通常オブジェクトを割り当てることで，通常オブジェクトとノイズオブジェクトを分離し，関係データからより適切にクラスタ構造を抽出する．本論文では，ドメイン 1 にのみノイズオブジェクトが存在するという仮定を置くが，複数のドメインにノイズオブジェクトの存在を仮定することも可能である．

提案モデルでは，ノイズオブジェクトの存在を仮定するドメイン 1 の各オブジェクトについて，潜在変数 $r_{1,i}$ と $y_{1,i}$ を導入する．変数 $r_{1,i}$ は SIRM と同様に，オブジェクト i がノイズオブジェクトかどうかを表し，オブジェクト i は，$r_{1,i}=1$ のとき通常オブジェクト，$r_{1,i}=0$ のときノイズオブジェクトとなる．また変数 $y_{1,i}$ はオブジェクト i に対するノイズクラスタのクラスタ割当であり，$r_{1,i}=1$ の場合，オブジェクト i は通常クラスタ $z_{1,i}$ に，$r_{1,i}=0$ の場合，オブジェクト i はノイズクラスタ $y_{1,i}$ に属しているとして扱う．提案モデルでは，$r_{1,i}$ の値によって，つまりオブジェクト i がノイズオブジェクトか否かによって，$x_{i,j}$ の生成過程が変化する．$r_{1,i}=1$ の場合，オブジェクト i,j はともに通常オブジェクトとなるため，IRM と同様に $x_{i,j}$ は関係確率 $\theta_{z_{1,i}\,z_{2,j}}$ をパラメータとする Bernoulli 分布から生成される．一方で，$r_{1,i}=0$ の場合，$x_{i,j}$ はノイズクラスタ $y_{1,i}$ が持つノイズ関係確率 $\phi_{y_{1,i}}$ をパラメータとする Bernoulli 分布から生成される．ここで，ϕ_m は，通常クラスタ間に定義されている関係確率 $\theta_{k,\ell}$ とは異なり，相手クラスタによらずノイズクラスタ m のみに依存する変数であることに注意する．これらを含めた

提案モデルにおける関係データの生成モデルは以下の通りである．

$$
\begin{aligned}
\lambda_1 &\sim \mathrm{Beta}(e, f) && 0 \leq \lambda_1 \leq 1, \\
r_{1,i} &\sim \mathrm{Bernoulli}(\lambda_1) && i = 1, ..., I,\ r_{1,i} \in \{0, 1\}, \\
z_{1,i} &\sim \mathrm{CRP}(\alpha_1) && i = 1, ..., I,\ z_{1,i} \in \{1, ..., \infty\}, \\
y_{1,i} &\sim \mathrm{CRP}(\beta) && i = 1, ..., I,\ y_{1,i} \in \{1, ..., \infty\}, \\
z_{2,j} &\sim \mathrm{CRP}(\alpha_2) && j = 1, ..., J,\ z_{2,j} \in \{1, ..., \infty\}, \\
\theta_{k,\ell} &\sim \mathrm{Beta}(a, b) && 0 \leq \theta_{k,\ell} \leq 1,\ k, \ell \in \{1, ..., \infty\}, \\
\phi_m &\sim \mathrm{Beta}(c, d) && 0 \leq \phi_m \leq 1,\ m \in \{1, ..., \infty\}, \\
x_{i,j} &\sim \mathrm{Bernoulli}(\theta_{z_{1,i},z_{2,j}}^{r_{1,i}} \phi_{y_{1,i}}^{1-r_{1,i}}) && i = 1, ..., I,\ j = 1, ..., J,\ x_{i,j} \in \{0, 1\}.
\end{aligned}
$$

ここで，λ_1 は SIRM と同様に，$r_{1,i}$ を生成する Bernoulli 分布のパラメータであり，ドメイン 1 のオブジェクト集合における通常オブジェクトの割合と解釈できる．また，$a, b, c, d, e, f, \alpha_1, \alpha_2, \beta$ はそれぞれハイパーパラメータであり，$Y = \{y_{1,i}\}_{i=1}^{I}$，$R = \{r_{1,i}\}_{i=1}^{I}$，$\Theta = \{\{\theta_{k,\ell}\}_{\ell=1}^{\infty}\}_{k=1}^{\infty}$，$\Phi = \{\phi_m\}_{m=1}^{\infty}$ である．提案モデルでは，CRP を通常クラスタ割当とノイズクラスタ割当の両方の生成過程において用いることで，通常クラスタとノイズクラスタの個数についての無限個の仮定を表現している．これにより提案モデルでは，多様なノイズが含まれる関係データに対して，与えられたデータから自動的にノイズクラスタの個数を推定し，多様なノイズオブジェクトに対してその性質に応じた異なるノイズクラスタを割り当てることができる．図 4 は提案モデルのグラフィカルモデルである．また，提案モデルにおける記号表記を表 1 にまとめる．

なお，本論文では記載しないが，ノイズオブジェクトが複数のドメインに存在すると仮定する場合においては，ノイズクラスタ間の関係確率を定義することでモデルを表現できる．また，上述の生成モデルは $x \in \{0, 1\}$ の場合であり，生成過程に用いる確率分布を変更することで多様な関係データに対して提案モデルは適用可能である．例えば，x が実数値をとる場合には Gaussian 分布，x が正の整数値をとる場合には Poisson 分布，x のとり得る値が有限個な場合には Categorical 分布をそれぞれ用いることでデータの生成過程を表現できる．

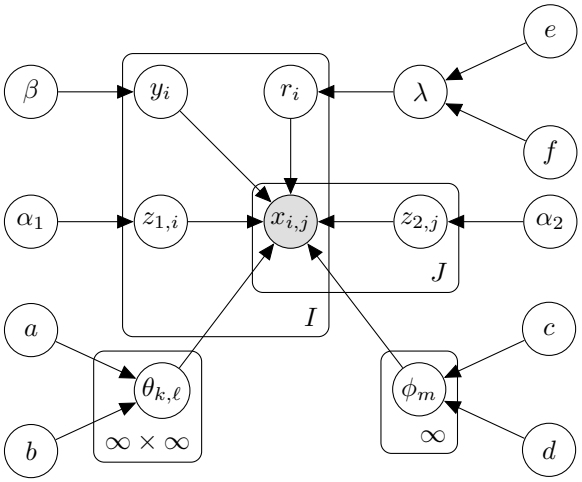


図 4: 提案モデルのグラフィカルモデル

4.2 先行手法との関係性

ノイズオブジェクトとノイズ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通常オブジェクトだと仮定する場合，提案モデルは IRM に一致する．具体的には，$\lambda_1 = 1$ のとき，提案モデルは IRM に対応する．また，オブジェクト集合を通常オブジェクトとノイズオブジェクトへと分割した上で通常オブジェクトのみを対象にクラスタリングを行う SIRM もまた提案モデルの特殊例とみなせる．SIRM では，ノイズオブジェクトはいずれも同じ関係確率で他のオブジェクトと関係を持つとして明示的にはクラスタリングの対象とされないものの，擬似的にはノイズオブジェクトを 1 つのノイズクラスタへと割り当てているとみなせる．つまり，$y_{1,i} = 1$ if $r_{1,i} = 0$ としてノイズクラスタの個数を 1 に限定する場合，提案モデルは SIRM と一致する．

図 5 は（a）IRM,（b）SIRM,（c）提案モデルのそれぞれにおいて仮定される関係データの潜在構造を表している．IRM では，関係確率が各クラスタの組み合わせについて個別に定義されるブロック構造のみを扱える．SIRM では，ノイズクラスタを導入することで相手のクラスタによらず関係確率が一定なクラスタを有する潜在構造を扱える．一方で，SIRM ではノイズクラスタを 1 つまでしか仮定できなかったのに対して，提案モデルでは任意の個数のノイズクラスタを持つ潜

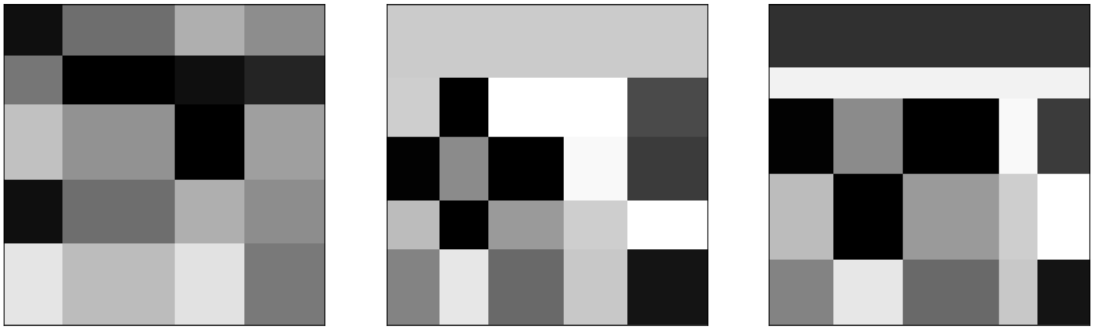
表 1: 記号表記

記号	説明
I	ドメイン 1 のオブジェクト総数
J	ドメイン 2 のオブジェクト総数
K	ドメイン 1 において現在割り当てられている通常クラスタの種類数
L	ドメイン 2 において現在割り当てられている通常クラスタの種類数
M	ドメイン 1 において現在割り当てられているノイズクラスタの種類数
$\theta_{k,\ell}$	通常クラスタ k, ℓ 間の関係確率
ϕ_m	ドメイン 1 のノイズクラスタ m が持つノイズ関係確率
λ_1	ドメイン 1 における通常オブジェクトの割合
$r_{1,i}$	ドメイン 1 のオブジェクト i がノイズオブジェクトかを表す変数
$z_{d,i}$	ドメイン d のオブジェクト i に割り当てる通常クラスタ
$y_{1,i}$	ドメイン 1 のオブジェクト i に割り当てるノイズクラスタ
$x_{i,j}$	オブジェクト i, j 間の関係の有無

在構造を扱える．この点において提案モデルは先行手法よりも適切に多様なノイズを含む関係データをクラスタリングできる．

4.3 推論

本節では，崩壊型ギブスサンプリングに基づいて，関係データ $\mathbf{X}$ から提案モデルの変数を推論する手続きについて説明する．提案モデルでは，共役性により，関係確率 Θ，ノイズ関係確率 Φ，通常オブジェクト割合 λ は解析的に積分消去が可能である．そのため，推論すべき変数は変数 R，ノイズクラスタ割当 Y，通常クラスタ割当 Z となる．提案法では，各オブジェクトについての変数を，その他の変数が与えられた状態における条件付き確率から同時にサンプリングし，それを全てのオブジェクトに対して繰り返すことで，変数を推論する．ここで，各ドメインにおける条件付き確率の式はノイズオブジェクトの存在を仮定するドメイン 1 と仮定しないドメイン 2 で異なることに注意する．



(a) IRM　　(b) SIRM　　(c) 提案モデル

図 5: 先行手法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

ドメイン 1 ではノイズオブジェクトの存在を仮定するため，各オブジェクト i について 3 つの潜在変数 $z_{1,i}$，$y_{1,i}$，$r_{1,i}$ を推論する必要がある．各変数が $z_{1,i} = k, y_{1,i} = m, r_{1,i} = n$ のように値を取るときの条件付き確率は次のようになる．

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

ここで $\setminus i$ はその値や集合がオブジェクト i を除いた場合のものであることを表す．上式の右辺の各項はそれぞれ次のように求まる．

$$
\frac{p(r_{1,i} = n, R^{\setminus i})}{p(R^{\setminus i})} = \begin{cases} e + \sum_{i' \neq i} r_{1,i'} & n = 1, \\ f + \sum_{i' \neq i} (1 - r_{1,i'}) & n = 0, \end{cases}
\tag{2}
$$

$$\frac{p(z_{1,i}=k, y_{1,i}=m, Y^{\setminus i}, Z^{\setminus i} |\, r_{1,i}=n, R^{\setminus i})}{p(Y^{\setminus i}, Z^{\setminus i} |\, R^{\setminus i})}$$
$$= \begin{cases} \dfrac{\beta}{\beta + \sum_{m'} V_{m'}^{\setminus i}} & m = M+1, \quad n = 0, \\ \dfrac{V_m^{\setminus i}}{\beta + \sum_{m'} V_{m'}^{\setminus i}} & m = 1, ..., M, \quad n = 0, \\ \dfrac{U_k^{\setminus i}}{\alpha_1 + \sum_{k'} U_{k'}^{\setminus i}} & k = 1, ..., K, \quad n = 1, \\ \dfrac{\alpha_1}{\alpha_1 + \sum_{k'} U_{k'}^{\setminus i}} & k = K+1, \quad n = 1, \end{cases} \tag{3}$$

$$\frac{p(\mathbf{X} |\, z_{1,i}=k, y_{1,i}=m, r_{1,i}=n, Z^{\setminus i}, R^{\setminus i})}{p(\mathbf{X}^{\setminus i} |\, Z^{\setminus i}, R^{\setminus i})}$$
$$= \begin{cases} \dfrac{B(c + S_m^+, d + S_m^-)}{B(c + S_m^{+\setminus i}, d + S_m^{-\setminus i})} & n = 0, \\ \displaystyle\prod_{\ell'} \frac{B(a + T_{k,\ell'}^+, b + T_{k,\ell'}^-)}{B(a + T_{k,\ell'}^{+\setminus i}, b + T_{k,\ell'}^{-\setminus i})} & n = 1, \end{cases} \tag{4}$$

where

$$U_k = \sum_i r_{1,i} \mathbb{I}(z_{1,i} = k),$$
$$V_m = \sum_i (1 - r_{1,i}) \mathbb{I}(y_{1,i} = m),$$
$$T_{k,\ell}^+ = \sum_i \sum_j r_{1,i} x_{i,j} \mathbb{I}(z_{1,i} = k) \mathbb{I}(z_{2,j} = \ell),$$
$$T_{k,\ell}^- = \sum_i \sum_j r_{1,i} (1 - x_{i,j}) \mathbb{I}(z_{1,i} = k) \mathbb{I}(z_{2,j} = \ell),$$
$$S_m^+ = \sum_i \sum_j (1 - r_{1,i}) x_{i,j} \mathbb{I}(y_{1,i} = m),$$
$$S_m^- = \sum_i \sum_j (1 - r_{1,i})(1 - x_{i,j}) \mathbb{I}(y_{1,i} = m).$$

ここで, K, M はそれぞれ各時点においてドメイン 1 のオブジェクトに割り当てられている通常クラスタとノイズクラスタの種類数であり，$k \in \{1, ..., K+1\}$，$m \in \{1, ..., M+1\}$，$n \in \{0, 1\}$ である．$B(\cdot)$ はベータ関数を表しており，$\mathbb{I}(\cdot)$ は a が真のときに $\mathbb{I}(a) = 1$ となり，偽のときに $\mathbb{I}(a) = 0$ となる指示関数である．なお，$z_{1,i}$，$y_{1,i}$，$r_{1,i}$ を同時にサンプリングするためには式 1 について k, m, n の取りうる値全てについて条件付き確率を計算する必要があるが，式 2，式 3，式 4 より n の値によって k か m の値のどちらかは計算上無視できるため，実際に計算する必要のある条件付き確率は $(K+1)+(M+1)$ 個となる．

ドメイン 2 のオブジェクトについては, 全てのオブジェクトが通常オブジェクトであると仮定するため，推論する変数は通常クラスタ割当 $z_{2,j}$ のみとなる．$z_{2,j} = \ell$ のときの条件付き確率は次のようになる．

$$\begin{aligned} p\left(z_{2,j} = \ell \mid \mathbf{X}, Y, Z^{\backslash j}, R\right) &= \frac{p\left(z_{2,j} = \ell, \mathbf{X}, Y, Z^{\backslash j}, R\right)}{p\left(\mathbf{X}, Y, Z^{\backslash j}, R\right)} \\ &\propto \frac{p(\mathbf{X}|\, z_{2,j} = \ell, Z^{\backslash j})}{p(\mathbf{X}^{\backslash j}|\, Z^{\backslash j})} \times \frac{p(z_j = \ell, Z^{\backslash j})}{p(Z^{\backslash j})}. \end{aligned} \tag{5}$$

右辺の各項はそれぞれ以下のように求まる．

$$\begin{aligned} &\frac{p(z_{2,j} = \ell, Z^{\backslash j})}{p(Z^{\backslash j})} \\ &= \begin{cases} \dfrac{\sum_{j' \neq j} \mathbb{I}(z_{2,j'} = \ell)}{\alpha_2 + \sum_{\ell'} \sum_{j' \neq j} \mathbb{I}(z_{2,j'} = \ell')} & \ell = 1, ..., L, \\ \dfrac{\alpha_2}{\alpha_2 + \sum_{\ell'} \sum_{j' \neq j} \mathbb{I}(z_{2,j'} = \ell')} & \ell = L+1, \end{cases} \end{aligned} \tag{6}$$

$$\frac{p(\mathbf{X}|\, z_{2,j} = \ell, Z^{\backslash j})}{p(\mathbf{X}^{\backslash j}|\, Z^{\backslash j})} \propto \prod_{k'} \frac{B(a + T^{+}_{k',\ell}, b + T^{-}_{k',\ell})}{B(a + T^{+\backslash j}_{k',\ell}, b + T^{-\backslash}_{k',\ell})}. \tag{7}$$

ここで, L は各時点においてドメイン 2 のオブジェクトに割り当てられている通常クラスタの種類数であり，$\ell \in \{1, ..., L+1\}$ である．これらの条件付き確率に従い全ての変数についてサンプリングを両ドメインで交互に繰り返すことで，Y, Z, R についての経験分布が事後分布の近似として得られる．

また，ハイパーパラメータについてもそれぞれ事前分布を置くことで事後分布からのサンプリングが行える．例えば，ハイパーパラメータ a をサンプリングす

Algorithm 1 提案モデルに対する推論アルゴリズム

initialize variables Y, Z, R and hyperparameters $a, b, c, d, e, f, \alpha_1, \alpha_2, \beta$
repeat
  for $i = 1$ to I **do**
    sample $y_{1,i}, z_{1,i}, r_{1,i}$ using (1)
  end for
  for $j = 1$ to J **do**
    sample $z_{2,j}$ using (5)
  end for
  sample hyperparameters $a, b, c, d, e, f, \alpha_1, \alpha_2, \beta$ using (8)
until end condition is met

るための条件付き確率は次のようになる．

$$
\begin{aligned}
&p(a = \hat{a}|X, Y, Z, R, b, c, d, e, f, \alpha_1, \alpha_2, \beta) \\
&\propto p(a = \hat{a})\, p(X, Y, Z, R|a = \hat{a}, b, c, d, e, f, \alpha_1, \alpha_2, \beta).
\end{aligned} \tag{8}
$$

本論文では，全てのハイパーパラメータの事前分布を，位置母数 0，尺度母数 2.5 の半コーシー分布として，各ハイパーパラメータごとに 10 の候補値を事前分布からサンプリングし，各値について計算される条件付き確率に基づいてハイパーパラメータをサンプリングする．

推論アルゴリズム全体の擬似コードを Algorithm1 に示す．提案モデルに対する推論手続きでは，最初にクラスタ割当 Y, Z と変数 R の初期化をしたのち，各ドメインでの全てのオブジェクトについての変数のサンプリングと，ハイパーパラメータのサンプリングを終了条件を満たすまで繰り返す．この手続きによりモデルの各変数に対する事後分布の近似が得られるが，実験では簡便のため，終了条件を満たした時点における変数のサンプル値を推定値として用いる．

ドメイン 1 における変数のサンプリングでは，各オブジェクトごとに $K+1$ 個の通常クラスタに対する時間計算量 $O(L)$ の割当確率の計算と $M+1$ 個のノイズクラスタに対する時間計算量 $O(1)$ の割当確率の計算が必要となる．また，ドメイン 2 における変数のサンプリングでは，各オブジェクトごとに $L+1$ 個の通常クラスタ

に対する時間計算量$O(K)$の割当確率の計算が必要となる．よって，全オブジェクトに対する変数のサンプリング1周あたりの時間計算量は$O((KL+M)I+KLJ)$となる．なお，同様の推論アルゴリズムをIRMに対して適用した場合，全オブジェクトに対する変数のサンプリング1周あたりの時間計算量は$O(KL(I+J))$となる．例えば，ドメイン1についてIRMが推定した通常クラスタの個数と，提案モデルが推定した通常クラスタの個数とノイズクラスタの個数の合計が同じである場合は，提案モデルについての推論アルゴリズムの方が時間計算量は少なくなる．

なお，上の推論手続きにおいては積分消去されていた$\theta_{k,\ell}$，ϕ_m，λの事後分布は，サンプリングで得られたY, Z, Rの値を用いると次のように書ける．

$$
\begin{aligned}
p(\theta_{k,\ell}|X,Y,Z,R) &= \mathrm{Beta}(a+T^{+}_{k,\ell}, b+T^{-}_{k,\ell}),\\
p(\phi_m|X,Y,Z,R) &= \mathrm{Beta}(c+S^{+}_{m}, d+S^{-}_{m}),\\
p(\lambda|X,Y,Z,R) &= \mathrm{Beta}(e+\sum_{i=1}^{I} r_{1,i}, f+\sum_{i+1}^{I}(1-r_{1,i})).
\end{aligned}
$$

実験ではこれらの分布における期待値を推定値として評価に用いる．

5. 実験

5.1 実験設定

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用性を示す．人工データとしては，2 ドメイン間の二値で表現される関係についての関係データを用いる．このデータには，ドメイン 1 に通常オブジェクトが 120，ノイズオブジェクトが 480 含まれており，ドメイン 2 には 800 の通常オブジェクトが含まれる．ドメイン 1 のクラスタ数は，通常クラスタが 6，ノイズクラスタが 4 であり，ドメイン 2 のクラスタ数は通常クラスタが 40 である．また，通常クラスタ間の関係確率はパラメータが共に 0.5 であるベータ分布からランダムに生成し，ノイズクラスタの関係確率はそれぞれ 0.15，0.30，0.45，0.60 とした．

実世界データとしては，20 Newsgroups [1]，MovieLens [2]，Google+ [22] の 3 つのデータセットを用いる．20 Newsgroups データセットは 20 のグループから投稿されたネットニュースについて約 400,000 文書を収集したものである．実験では，オリジナルのデータセットから 1000 文書を無作為抽出し，それらの文書において文書頻度上位 1000 個の単語を抽出する．抽出した文書と単語の各組についてその文書にその単語が出現したかどうかを二値の関係として構築した関係データを用いる．MovieLens のデータセットは，movielens という映画推薦のサービスにおける映画のレーティングを収集したもので，943 ユーザによる 1682 映画に対してのレーティングが含まれている．実験では，ユーザと映画の各組についてレーティングが存在するかを二値の関係として構築した関係データを用いる．Google+のデータセットは，Google+という SNS において特定のユーザを中心とした友人関係ネットワークを 132 個収集したものである．実験では 1 つのネットワークを抽出し，それに含まれる 527 ユーザ間のフォロー関係の有無を関係データとして用いる．なお，友人関係ネットワークはフォローと呼ばれる機能に基づく有向グラフとなっているため，一人のユーザをフォローする側とフォローされる側でそれぞれ別ドメインのオブジェクトとして扱う．さらに，Google+のデータセットから

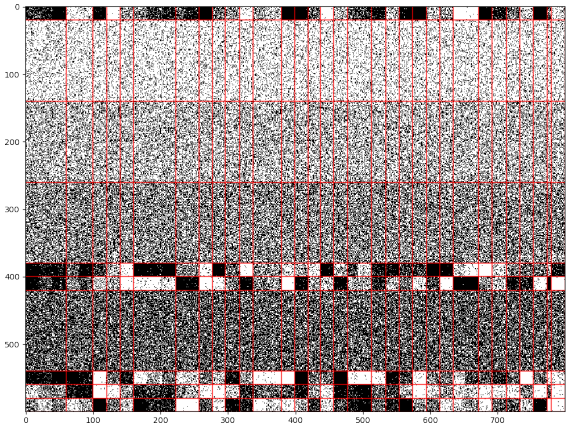
[1] http://qwone.com/~jason/20Newsgroups/
[2] https://grouplens.org/datasets/movielens/100k/

表 2: リンク予測における評価データに対する AUC

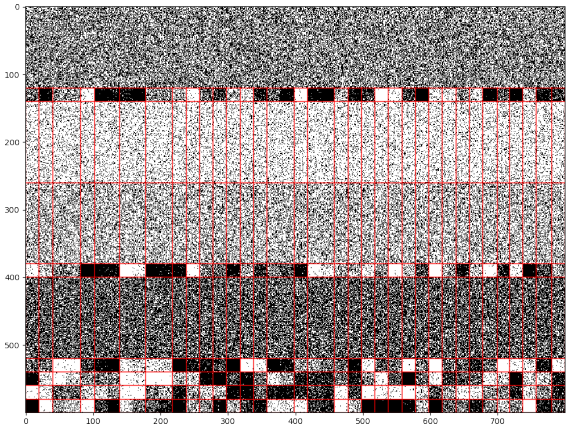
データセット	IRM	SIRM	提案法
人工データ	0.7585	0.7591	**0.7596**†
20News	0.8576	0.8578	**0.8592**†
MovieLens	0.9295	0.9305	**0.9310**
Google+	0.9601	**0.9603**	**0.9603**
Google+(w/noise)	0.9549	0.9554	**0.9561**†

は実世界データに人工的なノイズオブジェクトとして一定の確率で他のユーザをフォローするユーザを加えた関係データも構築する．具体的には，それぞれ 0.15，0.30，0.45，0.60 の関係確率を持つ 4 つのノイズクラスタからノイズオブジェクトとして 5 ユーザずつ生成し，計 20 ユーザを上述の Google+から構築した関係データにフォローする側のオブジェクトとして加えた新たな関係データを構築する．以下の実験結果ではこれらのデータセットをそれぞれ 20News，MovieLens，Google+，Google+(w/noise) と表記する．

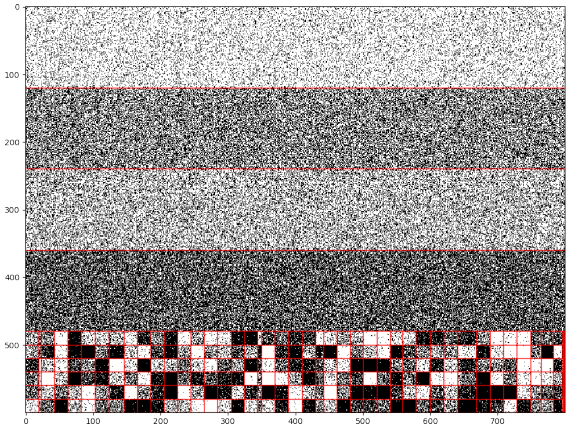
実験では，各データセットをそれぞれランダムに分割し，80%をモデルの変数推論に用いる訓練データ，残りの 20%を変数推論時には用いない評価データとする．比較手法には，提案モデルの基礎である IRM とノイズオブジェクトを扱えるモデルである SIRM を用いる．IRM と SIRM についても変数推論は提案法と同様の崩壊型ギブスサンプリングに基づく手法で行う．推論アルゴリズムは全変数についてのサンプリングを 500 周終えた時点で終了とする．また，変数推論は異なる初期状態から 10 試行し，最も訓練データに対する尤度が高くなった試行を評価に用いる．評価では，これらの手続きを 10 回繰り返し，各評価指標の平均をモデル間で比較する．評価は，リンク予測，クラスタリング，ノイズオブジェクト検出という 3 つのタスクについて行う．



(a) IRM

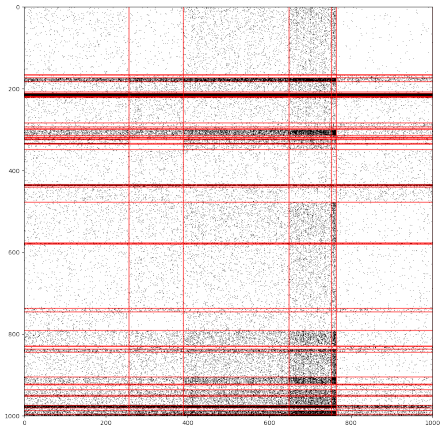


(b) SIRM

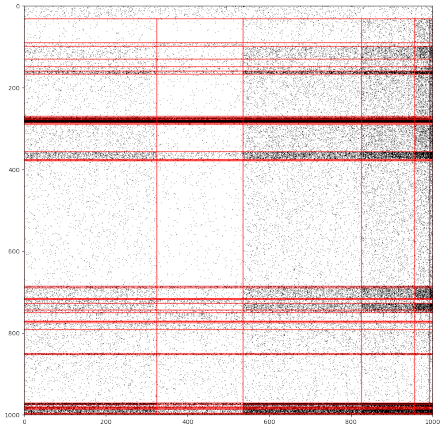


(c) 提案法

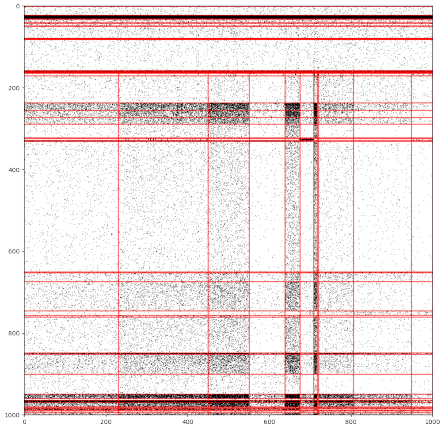
図 6: 人工データに対するクラスタリング結果



(a) IRM

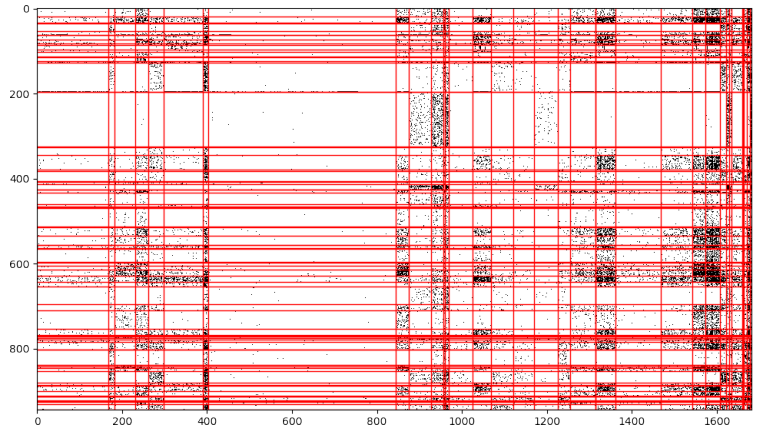


(b) SIRM

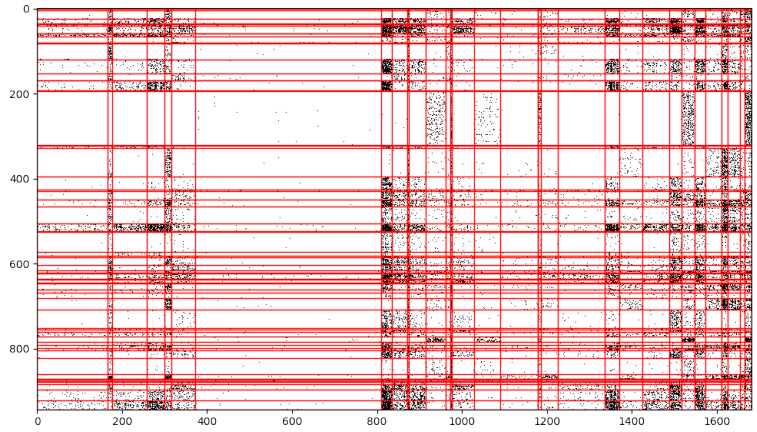


(c) 提案法

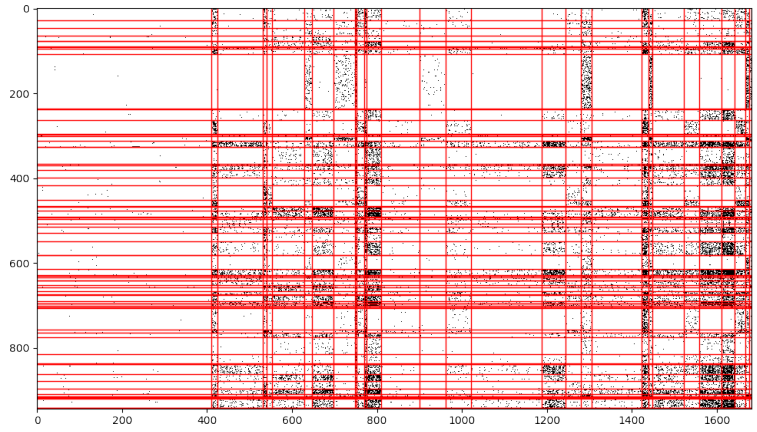
図 7: 20News データに対するクラスタリング結果（行：単語，列：文書）



(a) IRM

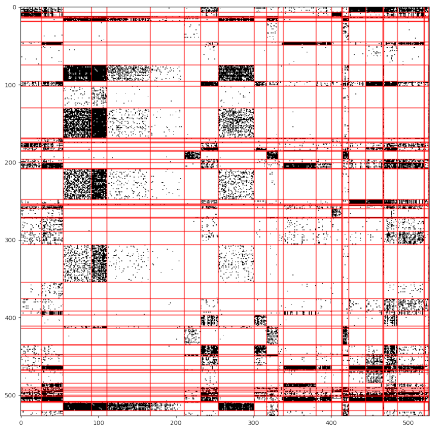


(b) SIRM

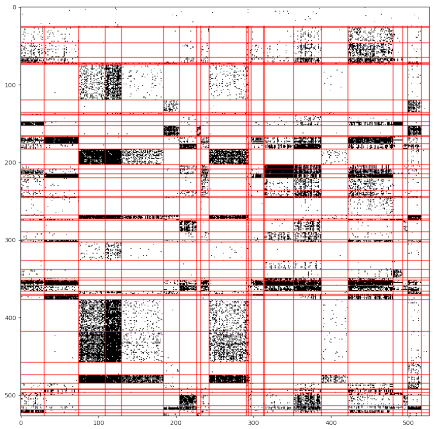


(c) 提案法

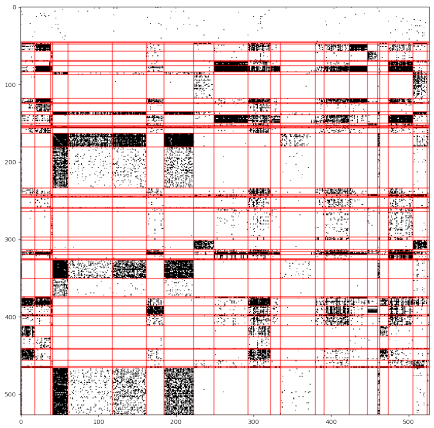
図 8: MovieLens データに対するクラスタリング結果（行：ユーザ，列：映画）



(a) IRM

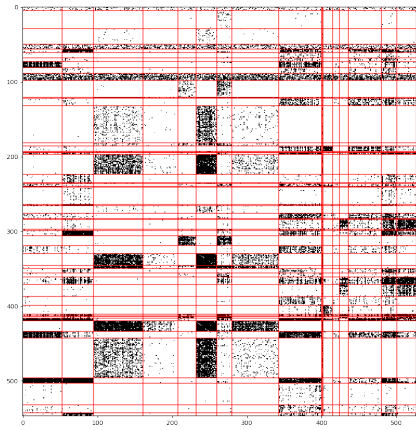


(b) SIRM

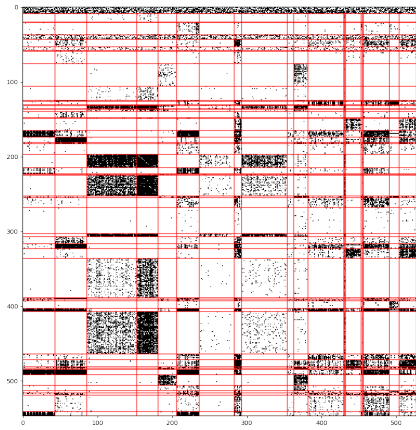


(c) 提案法

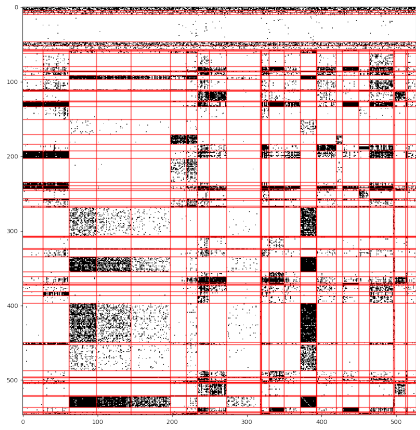
図 9: Google+データに対するクラスタリング結果（行 : フォローする側，列 : フォローされる側）



(a) IRM



(b) SIRM



(c) 提案法

図 10: Google+(w/noise) データに対するクラスタリング結果（行：フォローする側，列：フォローされる側）

表 3: 評価データに対する 1 要素あたりの対数尤度

データセット	IRM	SIRM	提案法
人工データ	-0.5640	-0.5628	**-0.5620**†
20News	-0.1966	-0.1963	**-0.1952**†
MovieLens	-0.1397	-0.1375	**-0.1373**
Google+	-0.1426	-0.1405	**-0.1402**
Google+(w/noise)	-0.1589	-0.1578	**-0.1566**

表 4: 推定されたクラスタ割当と真のクラスタ割当の ARI

データセット	IRM	SIRM	提案法
人工データ	0.9179	0.9320	**0.9502**†
Google+(w/noise)	0.6056	0.6140	**0.8314**†

5.2 実験結果

まず，それぞれのデータセットに対しての，IRM と SIRM と提案法によるクラスタリングの結果をそれぞれ図 6， 図 7， 図 8， 図 9，図 10 に例示する．各図は図 1 の（c）と同様に，行列表現された関係データをそれぞれの手法によるクラスタリングの結果推定されたオブジェクト割当に従って行と列を並び替えたものである．図中の赤い線はクラスタ間の境界を表しており，SIRM と提案法についての図の上部にある列方向に横切られていない赤い線はノイズクラスタとの境界を表している．人工データ，20News データ，Google+(w/noise) データについては，SIRM が一種類のノイズクラスタしか抽出できていないのに対して，提案法は複数のノイズクラスタを抽出できていた．この結果より，人工データはもちろん，20News データにも異なる関係確率を持つ複数のノイズオブジェクトが含まれていると考えられる．一方で，MovieLens のデータについては SIRM と提案法の両手法ともにノイズクラスタを抽出できず，Google+のデータについては両手法ともに 1 つのノイズクラスタしか抽出できなかった．この結果より，MovieLens データにはノイズオブジェクトがほとんど含まれておらず，Google+データには

表 5: ノイズオブジェクトの検出率

データセット	SIRM	提案法
人工データ	0.25	**1.00**†
Google+(w/noise)	0.45	**1.00**†

表 6: 人工データに対する性質別のノイズオブジェクト検出数

クラスタ (関係確率)	SIRM	提案法	正解
1 (0.15)	0.0	120.0	120.0
2 (0.30)	0.0	120.0	120.0
3 (0.45)	120.0	120.0	120.0
4 (0.60)	0.0	120.0	120.0

一種類のノイズオブジェクトしか含まれていないと考えられる．以下では，これらのデータについての考察に基づき，各タスクについての実験結果を議論する．

リンク予測では，関係データの一部が観測データとして与えられた場合における未観測データについての予測性能を評価する．評価指標には，ROC (Receiver Operating Characteristic) の AUC (Area Under Curve) と評価データに対する尤度を用いる．両指標とも値が大きいほど性能が高いことを示す．リンク予測の AUC を表 2 に，評価データ 1 要素あたりの対数尤度を表 3 に示す．表中の値は 10 回の実験における結果の平均値であり，太字の値は各行において最も高い値であることを示している．また，†のついた値は対応あり t 検定によって有意水準 5%で有意差があると判定された値である．なお，多重比較の補正は Holm の方法 [12] に基づいて行った．この表記法は，以下の表 4，表 5，表 6，表 7，表 9 でも同様である．両指標において，提案法は全てのデータセットについて最も高い値を達成した．また，人工データと 20News データについては両指標において，Google+(w/noise) データについては AUC の指標において，提案法が他の手法より有意に高い性能を発揮することも確認された．これらの結果は，提案法が多様なノイズが含まれる関係データにおける未観測データの予測に役立つことを表している．一方で，MovieLens データと Google+データについての実験では提案法と比較手法の性能

表 7: Google+(w/noise) データに対する性質別のノイズオブジェクト検出数

クラスタ (関係確率)	SIRM	提案法	正解
1 (0.15)	0.0	5.0	5.0
2 (0.30)	0.5	5.0	5.0
3 (0.45)	4.0	5.0	5.0
4 (0.60)	4.5	5.0	5.0

表 8: 提案法により 20News データから検出されたノイズクラスタ例

クラスタ	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

に有意差は認められなかった．これは，これらのデータにはノイズオブジェクトがほとんど含まれていない，あるいは一種類のノイズオブジェクトしか含まれていなかったためと考えられる．この結果より，ノイズオブジェクトが含まれない，あるいは含まれるノイズオブジェクトの性質が一様な関係データに対しては，提案法は従来法と同様の挙動をすることが示唆される．

クラスタリングでは，真のクラスタ割当と各手法によるクラスタリングの結果がどの程度一致するかを評価する．このタスクでは，真のクラスタ割当がわかっている人工データと Google+(w/noise) データにおける人工的なノイズオブジェクトのみを評価に用いる．評価指標には，ARI (adjusted Rand index) [13] を用いる．ARI は二つのクラスタリングの類似度を測る指標であり，各手法によって推定されたクラスタ割当と真のクラスタ割当の ARI を求めることで各手法によるクラスタリングの性能を評価できる．ARI は-1 から 1 の値を取り，実験では 1 に近い値ほど性能が高いことを示し，0 に近い値ほどランダムなクラスタリングであるこ

表 9: 計算時間 (秒)

データセット	IRM	SIRM	提案法
人工データ	**1080**	1212	1344
20News	**1557**	1732	2252
MovieLens	9279	**8890**	9128
Google+	2027	1964	**1872**
Google+(w/noise)	**1835**	1886	2040

とを示す．表 4 に推定されたクラスタ割当と真のクラスタ割当の ARI を示す．提案法は両データセットにおいて最も高い値を達成し，比較手法と有意な性能差があることが確認された．この結果は，提案法がノイズオブジェクトに悪影響を受けることなく多様なノイズが含まれる関係データを適切にクラスタリングできることを表している．

ノイズオブジェクトの検出では，関係データに含まれるノイズオブジェクトをノイズオブジェクトとして推定できるかを評価する．このタスクでも，真のクラスタ割当がわかっている人工データと Google+(w/noise) データにおける人工的なノイズオブジェクトのみを評価に用いる．また，ノイズオブジェクトを明示的に推定できない IRM は評価しない．表 5 に，各手法によって真のクラスタ割当がノイズクラスタであるオブジェクトの内ノイズクラスタに割り当てることができたものの個数の平均を示す．また表 6，7 にはノイズオブジェクトの真のクラスタ割当別に，すなわちノイズオブジェクトの性質別に検出できたノイズオブジェクトの個数を示す．SIRM はほとんどのノイズオブジェクトの検出に失敗した一方で，提案法は正確にノイズオブジェクトを検出できた．また，SIRM は検出できるノイズオブジェクトの性質に偏りがあった一方で，提案法は関係確率の値の大小に依らずノイズオブジェクトを検出できた．人工的なノイズオブジェクトについての評価ではあるものの，この結果より提案法は従来法より関係データに含まれるノイズオブジェクトの検出に有用であると考えられる．

表 8 に，提案法が 20News データから抽出した単語についてのノイズクラスタの例を示す．ノイズクラスタ 1 の単語としては，ネットニュースとして配信され

る 20News のどの文書にもヘッダとして出現する単語のクラスタが抽出できていた．またノイズクラスタ 2 には曜日を表す単語等が，ノイズクラスタ 3 には記号文字等が含まれていた．これらの単語はいずれも文書の種類によらず出現する単語である一方で，その性質はノイズクラスタ 1 とノイズクラスタ 2 と 3 ではそれぞれ異なると考えられる．この結果からも，提案法はねらい通りに多様なノイズオブジェクトを扱えることが示唆される．

表 9 に，各モデルに対する推論アルゴリズムの計算時間を示す．なお，実験に用いた計算機の仕様は，CPU：Intel(R) Xeon Gold 6130，RAM：256GB であり，各手法とも実装には Python を用いた．リンク予測において提案モデルが高い性能を示した，人工データ，20News，Google+(w/noise) のデータセットについては提案モデルが最も長い計算時間を要した．これは，多様なノイズを含むデータについては，提案モデルは IRM や SIRM より詳細なクラスタ構造を抽出できたために，サンプリングにおいて計算する条件付き確率の数が多くなってしまったからだと考えられる．一方で，リンク予測においてモデル間で性能の差がほとんどなかった MovieLens と Google+のデータセットについては，計算時間についてもモデル間の差が小さかった．これは，ノイズをほとんど含まないデータについては，提案モデルと IRM や SIRM で抽出するクラスタ構造に大きな差がなく，サンプリングにおいて計算する条件付き確率の数が同等になったためだと考えられる．

6. おわりに

本論文では，関係データに対する頑健なクラスタリング手法として，複数の異なる性質を持つノイズオブジェクトを含む関係データから潜在的なクラスタ構造を抽出できる確率モデルを提案した．提案モデルと既存手法の関係については，IRM や SIRM が提案モデルの特殊例とみなせることを確認した．また，提案モデルに対する推論アルゴリズムとして崩壊型ギブスサンプリングに基づく効率的なアルゴリズムを導出した．実験では，リンク予測，クラスタリング，ノイズオブジェクトの検出という 3 つのタスクについて評価した．人工データを用いた実験では，提案法が従来の手法よりも優位に高い性能を示し，多様なノイズを含む関係データに対する提案法の有用性が確認された．また，実世界の関係データを用いた実験では，提案法は従来法と同等以上の性能を示し，実世界データに対する提案法の有用性も確認された．今後の展望としては，1 つのオブジェクトに複数のクラスタを割り当てる関係データクラスタリング手法に対する無限個のノイズクラスタという機構の適用や，教師なしクラスタマッチング [16] 等の他のタスクへの提案法の適用を考えている．また，関係データに含まれない周辺情報の利用 [27]，動的なデータへの対応 [14] といったモデルの拡張や，分散処理に対応する高速な推論アルゴリズムの構築も展望している．

謝辞

本研究を進めるにあたり，多くの方々からご協力とご支援を頂いたことを心より感謝いたします．

コミュニケーション学研究室の澤田宏教授と岩田具治准教授には二年間直接的に研究を指導いただきました．多忙な中，週ごとの研究進捗に対して議論やコメントをいただいただけでなく，私の拙い論文や発表資料を辛抱強く添削いただき，大変励みになりました．心より感謝いたします．

主指導教員の松本裕治教授と副指導教員の新保仁准教授には，定例の研究会や勉強会を通して研究に対する有益な議論やコメントをいただきました．また，連携講座の学生である私に対しても，研究室の他の学生と遜色ない充実した研究環境を与えていただきました．本当にありがとうございました．

副指導教員の中村哲教授には本研究の発表の場において貴重なコメントをいただきました．ここに感謝いたします．

また，NTT コミュニケーション基礎科学研究所の協創情報研究部の皆様と自然言語処理学のスタッフの方々には，研究生活を送る上で様々な支援やアドバイスをいただきました．ここに感謝いたします．

最後に，二年間の学生生活を共に過ごした，コミュニケーション学研究室と自然言語処理学の学生の皆様に感謝いたします．

参考文献

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[2] David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

[3] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems*, pages 241–248, 2007.

[4] Rajesh N Davé and Sumit Sen. Robust fuzzy clustering of relational data. *IEEE Transactions on Fuzzy Systems*, 10(6):713–727, 2002.

[5] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274. ACM, 2001.

[6] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[7] Yue Guan, Michael I Jordan, and Jennifer G Dy. A unified probabilistic model for global and local unsupervised feature selection. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1073–1080, 2011.

[8] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

[9] Richard J Hathaway and James C Bezdek. Nerf c-means: Non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27(3):429–437, 1994.

[10] Richard J Hathaway, John W Davenport, and James C Bezdek. Relational duals of the c-means clustering algorithms. *Pattern Recognition*, 22(2):205–212, 1989.

[11] Peter D Hoff. Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics*, 61(4):1027–1036, 2005.

[12] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[13] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[14] Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, and Joshua B Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. In *Advances in Neural Information Processing Systems*, pages 919–927, 2010.

[15] Katsuhiko Ishiguro, Naonori Ueda, and Hiroshi Sawada. Subset infinite relational models. In *Artificial Intelligence and Statistics*, pages 547–555, 2012.

[16] Tomoharu Iwata and Katsuhiko Ishiguro. Robust unsupervised cluster matching for network data. *Data Mining and Knowledge Discovery*, 31(4):1132–1154, 2017.

[17] Tomoharu Iwata, James Robert Lloyd, and Zoubin Ghahramani. Unsupervised many-to-many object matching for relational data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):607–617, 2016.

[18] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Modeling social annotation data with content relevance using a topic model. In *Advances in Neural Information Processing Systems*, pages 835–843, 2009.

[19] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[20] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

[21] Bo Long, Zhongfei Mark Zhang, Xiaoyun Wu, and Philip S Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 585–592. ACM, 2006.

[22] Julian McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, pages 539–547, USA, 2012. Curran Associates Inc.

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

[24] Morten Mørup and Mikkel N Schmidt. Bayesian community detection. *Neural Computation*, 24(9):2434–2456, 2012.

[25] Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[26] Iku Ohama, Hiromi Iida, Takuya Kida, and Hiroki Arimura. An extension of the infinite relational model incorporating interaction between objects. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 147–159. Springer, 2013.

[27] Masafumi Oyamada and Shinji Nakadai. Relational mixture of experts: Explainable demographics prediction with behavioral data. In *2017 IEEE International Conference on Data Mining*, pages 357–366. IEEE, 2017.

[28] Konstantina Palla, David A Knowles, and Zoubin Ghahramani. An infinite latent attribute model for network data. In *Proceedings of the 29th International*

Coference on Machine Learning, pages 395–402. Omnipress, 2012.

[29] Mikkel N Schmidt and Morten Morup. Nonparametric bayesian modeling of complex networks: An introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, 2013.

[30] Ilya Sutskever, Joshua B Tenenbaum, and Ruslan R Salakhutdinov. Modelling relational data using bayesian clustered tensor factorization. In *Advances in Neural Information Processing Systems*, pages 1821–1828, 2009.

[31] Joyce Jiyoung Whang, Piyush Rai, and Inderjit S Dhillon. Stochastic blockmodel with cluster overlap, relevance selection, and similarity-based smoothing. In *2013 IEEE International Conference on Data Mining*, pages 817–826. IEEE, 2013.

関連発表リスト

国内会議（査読なし・口頭発表）

- 武田 悠佑, 岩田 具治, 澤田 宏, “多様なノイズに頑健な関係データのクラスタリング,” 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), 福井, Mar. 2018.

その他の発表

国内会議（査読なし・ポスター発表）

- 武田 悠佑, 岩田 具治, 澤田 宏, “Attention 機構を用いた集合データに対する確率的ニューラル回帰モデル,” 第 21 回情報論的学習理論ワークショップ (IBIS2018), 北海道, Nov. 2018.