学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　大濵　郁

学　位　論　文　題　名

Relevance-dependent Biclustering Models for Relational Data Analysis
（関係データ分析のための関連度依存型共クラスタリングモデルの研究）

昨今のインターネットの発展により, 人の行動を記録した膨大なデータの活用による社会の更なる発展が望まれている. 特に, 産業界においては, 購買履歴やソーシャルネット上の交友関係など, 消費者行動を表すデータの分析によるビジネス改善に大きな期待が寄せられている. 前述した人やモノの結びつき (リンク) を表す二値行列型のデータは関係データと呼ばれる. 一般に, このような関係データには, 消費者の嗜好や商品の性質を表す有益な構造情報が潜在している. 従って, 関係データから潜在構造を発見する技術の確立は, 学術と産業の両方の観点で重要な課題と言える.

関係データの潜在構造を抽出する代表的手法である共クラスタリングは, 関係データの行と列を同時クラスタリングすることでデータの要約であるブロック構造を得る技術であり, 近年では統計モデルに基づく共クラスタリングの研究が本格化してきた. Kemp らが提案した無限関係モデル (Infinite Relational Model, IRM) は,「各観測値は, 自身が所属するブロック固有の分布から生成された」という仮定に基づき, クラスタ割当ての事後分布を学習する. IRM はクラスタ数をデータから自動的に学習することができる. さらに, IRM は周辺化ギブスサンプリングにより効率的に大域最適解を学習可能である.

しかし, IRM に代表される標準的な共クラスタリングモデルには大きな欠点がある. これらの共クラスタリングモデルでは, 個々のブロックは一様な確率密度を持つという仮定が置かれている. しかし, 実世界においてこの仮定は不適切であることが多い. 例えば, E コマースサイトの購買履歴において, 購買行動は, 消費者の好みや商品の特徴だけではなく, 予算状況や商品の知名度など, 本質的でない要因にも影響される. 従って, 実世界の関係データは, 理想的なブロック構造を表す分布に対して, 各オブジェクト固有の事情により歪められた分布から生成されたと考えるべきである. このような状況で共クラスタ構造を捉えるためには, オブジェクト毎にリンク確率に作用する新しい潜在因子を考える必要がある.

本研究では, 関連度依存型共クラスタリングという新しい共クラスタリング問題を議論する. 関連度依存型共クラスタリングでは, クラスタ割当てとは別に, 各オブジェクトが理想的なブロック構造に従う強さを表す「関連度」という潜在変数を考える. つまり, 関連度が大きいオブジェクトはブロック構造の分布に強く従い, 関連度が小さいオブジェクトはブロック構造への従属度が弱いことを意味する. 従って, 関連度依存型共クラスタリングにより得られるブロックは非一様な密度を持ち, その非一様性を関連度の効果として説明することができる. さらに, 関連度依存型共クラスタリングでは, 高い関連度を持つ少数のオブジェクトを調べるだけで, そのクラスタの意味を理解することが出来るという実用上の利点を有する. 本研究は, 関連度依存型共クラスタリングのための統計モデルとその効率的学習方法の開拓を目的とする.

まず第 3 章では, 論理関数を用いた関連度依存性のモデル化を提案する. 具体的には, 関係データの各エントリに対して, そのエントリがブロック構造を表す分布 (前景分布) から生成されたのか,

ブロック構造とは無関係な分布 (背景分布) から生成されたのかを選択する二値潜在変数を導入する. そして, その二値潜在変数の値が, 関係する行オブジェクトと列オブジェクトの Bernoulli 試行の論理演算で決定される仕組みを導入する. ここで, 各オブジェクトの Bernoulli 試行を制御する確率が関連度を表すパラメータとみなすことができる. 前述の分布選択機構を用いて IRM を拡張した Relevance-dependent IRM (RDIRM) を提案する. RDIRM は, IRM と同様に, クラスタ数を自動的に推定することが出来る. さらに, RDIRM は, 分布の共役性により, 周辺化ギブスサンプリングによる効率的な学習が可能である.

次に第 4 章では, 前述した論理関数によるアプローチの一般化について議論する. RDIRM には 2 つの欠点がある. まず, 行と列の Bernoulli 試行を論理演算する相互作用関数を人手で決める必要がある. しかし, 最適な関数形状は自明ではないため, 関数の形もデータから学習できることが望ましい. さらに, RDIRM は前景分布と背景分布の混合によりリンク確率を調節するため, リンク確率は 2 つの分布の混合で定義される範囲でしか調節されない. しかし, 実際の関係データにはスパムのようにリンク確率を増加させるオブジェクトも含まれる. 従って, より自由に確率調節が可能な関連度依存性モデルを考える必要ある. そこで, RDIRM の論理関数を混合メンバシップモデルへ一般化することで, この 2 つの課題を解決する. 具体的には, RDIRM の相互作用関数を連続緩和して, 全ての論理関数を含むように拡張する. さらに, 関連度を多次元の確率ベクトルに拡張することで, 3 つ以上の分布を混合できる分布選択機構を提案する. そして, 前述の分布選択機構を組み込んだ Multi-Layered IRM (MLIRM) を提案する. MLIRM はスパム的にリンクを張るような現象も関連度の影響として説明することができる. また, MLIRM は RDIRM と同様に周辺化ギブスサンプリングにより効率的に学習することが可能である.

最後に第 5 章では, リンク関数を用いた関連度依存性モデリングの方法を議論する. RDIRM や MLIRM では, 分布を選択するための潜在変数が存在するため, 行列サイズオーダーの計算量を要するという欠点がある. さらに, MLIRM では, 獲得された相互作用関数の解釈が困難になるという欠点がある. これらの欠点を解決するために, Bernoulli-Poisson (BerPo) リンク関数という特殊な関数を用いた関連度依存性モデリングを考える. BerPo リンク関数は, 非負値の量を確率に変換するリンク関数である. そこで, 各オブジェクトの関連度とブロックの典型的リンク強度を非負の値として定義する. そして, それら非負値の積を BerPo リンク関数により確率に変換して関連度依存型の二値行列を生成する関連度依存型 Bernoulli 分布 (R-BD) を導入する. R-BD では, 背景分布を導入する必要がないため, 分布選択のための潜在変数が不要である. また, MLIRM のように関連度の多次元化や相互作用関数の複雑化をさせることなく, 確率を 0.0 から 1.0 まで自在に調節できるため, 関連度の効果を容易に解釈できるという利点がある. さらに, R-BD のパラメータは全て周辺化消去することが出来るため, 極めて効率的に学習可能である. 本研究では, この R-BD を各ブロックの要素分布として組み込んだ Relevance-dependent Infinite Biclustering (R-IB) を提案する. R-IB は, 周辺化ギブスサンプリングにより, RDIRM や MLIRM だけでなく IRM よりも高速に学習することができる.

本研究を通して, 関連度依存型共クラスタリングのための複数の統計モデルを開拓した. 本研究で導入した関連度のモデル化方法とその効率的な学習方法は, 共クラスタリングだけでなく, 行列分解などより一般的な機械学習問題に対しても適用可能なアイデアであり, 今後の更なる展開が期待できる.