購買行動分析のためのベイジアンネットワークの構造学習に関する研究

1X13C032-1　河部 瞭太
指導教員　後藤 正幸

1　研究背景と目的

変数間の因果関係を表現するモデルとしてベイジアンネットワーク（以下，BN）がある．BN は，因果関係があるノード (i.e. 確率変数) 間を有向リンクで繋ぎ，非循環有向グラフの形でネットワークを学習し，学習した確率構造を用いて一部のノードの観測値が与えられたもとで未観測のノードの事後確率を予測するモデルである．BN は，幅広い問題に適用可能であり，多くのデータ解析へ適用されている．本研究では，商品の購買履歴データに適用し，商品購買の共起の構造を可視化する方法としての BN モデルを研究対象とする．

一般に，小売業における商品は大分類，中分類，小分類のように階層的な分類がされている．しかし BN では，変数の階層構造が想定されていないため，1 つの分類に着目した BN を用いてモデル化をしなければならない．顧客の購買行動を分析するに際して大分類や中分類に着目した場合，カテゴリが大まかなため大局的観点からの分析となり，顧客の個々の商品に対する購買行動を詳細に表現することは難しい．一方，小分類のみに着目してネットワークを作成した場合，ノード数が膨大となり，ネットワークが過度に複雑化して全体の関係性の把握が困難になるという問題がある．従って商品の階層構造を利用し，大局的観点，局所的観点を同時に表現できる新たな BN のモデル化とその学習法が望まれる．

一方，BN は構造学習の方法によって予測・推論の精度が左右される．BN の構造を学習するための手法として，計算時間を削減しつつ精度の高い構造を学習する The max-min hill-climbing 法（以下，MMHC）[1] が Tsamardinos らによって提案されている．この方法では，まずノード間を独立検定し，リンクが引かれる可能性が高いと思われる 2 つのノード間に対して無向グラフを作成し，その後スコアベースの手法を用い探索的に有向リンクを作成する．

本研究では変数の階層構造を考慮した BN モデル及び MMHC を用いた学習法を提案する．本手法は商品の階層性に着目し，上位層のネットワークでの因果関係を考慮しながら下位層のネットワークを構築する事を可能とする．また，株式会社マクロミル提供の階層性を持つ消費者購買履歴データに提案手法を適用し，その分析結果を示す．

2　従来手法

2.1　ベイジアンネットワーク

BN は因果関係を確率的に表現するモデルである．本研究では，BN の各ノードを各商品購買の事象とし，入力値として，商品の購買の頻度に対してカテゴリデータ化した値を用いることにする．BN を対象問題に適用する際には，BN の構造を学習するステップと学習した構造を用いて確率を推論する 2 つのステップがある．まず，構造学習では学習データの各ノード間の因果関係を有向リンクで表し，ネットワークを作成する．ここで有向リンクとは因果関係のある 2 つのノード間に引かれ，条件付き確率の向きを規定したものである．次に確率推論では観測値が得られたノードに値を入力し，未観測のノードの事象の発生確率を連鎖的に予測する．

2.2　The max-min hill-climbing 法

BN における構造学習の一手法として Tsamardinos らによって提案された MMHC[1] がある．MMHC ではリンクを引くことができる有向リンク候補を作成し，その制約のもと，スコアベースの手法を用いてネットワークを学習することで，計算時間を削減することを可能としている．

2.2.1　有向リンク候補の作成

各ノード間の有向リンク候補を作成するため，尤度比検定の一種である G^2 検定を用い，ノード間の独立性の検定を行う．T を対象ノード，X を対象ノードとの検証をするノード，Z を既に関連があるとされたノード集合，a，b は各ノード X，T の取る有限個の値，c はノード集合 Z の取る値の有限個の組合わせとする．$S_{X,\,T,\,Z}$ をデータにおけるX，T，Zの組み合わせの発生頻度とすると,G^2 検定の統計量は以下の (1) 式で与えられる．

$$G^2 = 2\sum_{a,b,c} S_{X=a,T=b,Z=c} \ln \frac{S_{X=a,T=b,Z=c} S_{Z=c}}{S_{X=a,Z=c} S_{T=b,Z=c}} \qquad (1)$$

2.2.2　有向リンク構造の作成

BN の構造学習の手法は主に制約ベースの手法とスコアベースの手法に大別される．ここでスコアベースの手法の 1 つとして Greedy Hill-Climbing アルゴリズム（以下，GHC）に着目する．GHC ではリンクが存在しない状況を初期状態とし，有向リンクを 1 つ「追加する」,「反転する」,「削除する」のいずれかの操作により Bayesian Information Criterion（以下，BIC）が改善するように構造を学習していく．BIC はモデル選択のために使用される指標であり，尤度関数に対しモデルの複雑さとデータサイズからなる罰則項を取り入れた基準である．リンク構造の作成のステップでは 2.2.1 項で作成した無向グラフのみに制限して GHC を用いることで，効率的なネットワーク構築を可能とする．

3　提案手法

BN では，データの階層構造が想定されていないため，上位層で学習したネットワークにおける因果関係の方向と下位層で学習したネットワークにおける因果関係の方向が矛盾する可能性がある．また，MMHC を用いた BN 学習では商品の小分類データのようなノード数の多いデータを用いるとネットワークが過度に複雑になり解釈が困難になるという問題が存在する．そこで，本研究では上位層のネットワークにおいて共起関係があった 2 つの上位層ノード間のみを抽出して，抽出したペアに対して下位層レベルでネットワークを構築するという新たな BN の構造学習法を提案する．このモデルでは上位層レベルのネットワーク構造に存在した共起関係の方向にのみリンクを引く場所及び方向を制限して，下位層に対し MMHC を用いた構造学習を行う．これによりノード数が多い状況でもノード間の共起に関する解釈を行う事，上位層での共起を下位層レベルのノード間の共起関係に反映させる事が可能となる．以下に提案アルゴリズムを示す．

Step1) 上位層に着目した購買履歴データに MMHC を適用し，上位層レベルでのネットワークを構築する．

Step2) Step1 で共起関係が確認された 2 つの上位層のペアをリンクの個数分抽出する．

Step3) Step2 で抽出したペアそれぞれに対し上位層レベルでの共起関係の方向全てを有向リンク候補とし，更に MMHC で下位層レベルでのネットワークを構築する．

4　実験

4.1　実験条件

本研究では提案モデルの有用性を検討するために株式会社マクロミル提供の消費者購買履歴データ QPR（Quick Purchase Report) を使用した．データ内容は首都圏のサイト利用者の 2015 年 1 年間の購買データであり，顧客それぞれが

購買の度にバーコードリーダーを使用して購買データをデータベースへと送信した情報が蓄積されている．商品の中分類は 25 種類，小分類は 275 種類であった．また，利用者数は 7,872 人, 購買データ数は 7,827,088 件であった．

ここで，BN はカテゴリカルな変数に対して適用されるモデルであるため，顧客ごとの中，小分類の各商品における購買点数データに対して，各商品の購買点数を以下の 3 つの場合分けでカテゴリデータ化する．

Case1) ある商品の購買点数の中央値が 0 個であるとき，購買点数が「0」と「1 以上」で 2 値化する．

Case2) ある商品の購買点数の中央値が 1 個であるとき，購買点数が「1 以下」と「2 以上」で 2 値化する．

Case3) Step1，2 に当てはまらなかった場合，所属人数が均等になるように 3 値化する．

4.2 中分類データに着目した分析

まず，提案手法の Step1 によって得られた中分類レベルでのネットワークを図 1 に示す．ここでは，紙面の都合で一部を取り上げている．

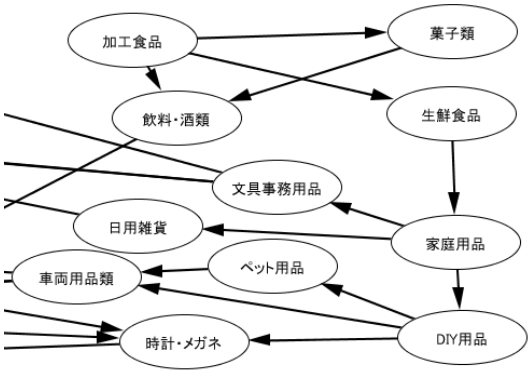

図 1. 中分類レベルでのネットワークの一部

図 1 より，例えば「生鮮食品」と「家庭用品」の購買に共起関係があることがわかる．実際に条件付き確率を計算すると,「家庭用品」をよく買う確率は,「生鮮食品」をあまり買わない場合は 3.8%であるのに対し,「生鮮食品」をよく買う場合は 56.4%と大きく向上している．この 2 つの中分類間について小分類レベルで分析した結果を以下の 4.3 節に示す．

4.3 小分類データに着目した分析

中分類データでの共起関係を保持しながら求めた提案手法の Step3 で得られた小分類レベルでのネットワークの一例を図 2 に示す．

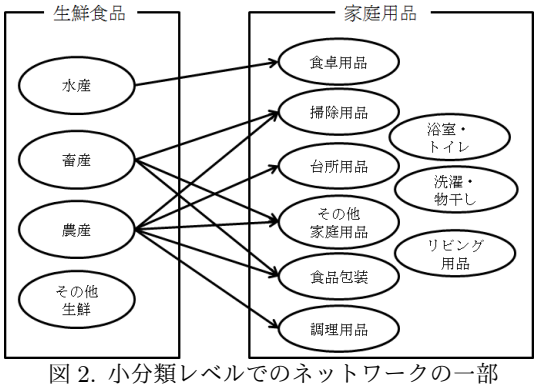

図 2. 小分類レベルでのネットワークの一部

図 2 より，例えば「水産」と「食卓用品」の購買には共起関係があることがわかる．4.2 節と同様に条件付き確率を計算すると「食卓用品」をよく買う確率は,「水産」をあまり買わない場合は 42.4%であるのに対し,「水産」をよく買う場合は 72.4%と大きく向上している．このような 2 つの中分類間に限定した小分類レベルでのネットワークを全ての中分類のペアで作成し，繋げることで中分類の共起関係を保持した状態で小分類全体のネットワークを構築することができる．

4.4 シミュレーション分析

ここでは，本実験のモデルを用いてどのように顧客の購買の向上施策を立案できるかを示すための活用例として，シミュレーション分析の結果を示す．どの商品の購買が他の商品の購買へと強く繋がるかを求め，全体の購買の確率を最も高める商品カテゴリを発見する．そのため，実験で学習した商品の各階層でのネットワークを用いて，対象ノードにおける 4.1 節で示した購買カテゴリの中で，購買点数最大の購買カテゴリが生起したと仮定したときの他ノードに対する影響について確率推論を行う．具体的には，対象ノードの購買点数最大の購買カテゴリの生起確率を 1 に変化させた際の，対象ノード以外のノードにおける購買点数最大の購買カテゴリの生起確率の平均変動率を求める．これにより，ネットワーク全体での購買増へのインパクトを示すことが出来る．即ち平均変動率 $R_l(0 \leqq R_l \leqq 1)$ は，ある商品の購買による他の商品の購買への繋がりやすさを示す値である．

ここで $P\,(n_{il}^a)$ を l 番目のノード n_l の観測値が購買点数最大としたときに予測される i 番目のノードの最も購買の多い購買カテゴリの生起確率，$P\,(n_i^p)$ を i 番目のノードの最も購買の多い購買カテゴリの事前確率，D をノードの総数とし，全ノードに対する平均変動率 R_l を以下の式 (2) で定義する．

$$R_l = \frac{\Sigma_{i \neq l}(P\,(n_{il}^a) - P\,(n_i^p))}{D-1} \tag{2}$$

D 個すべての商品の中分類に対して (2) 式の R_l を求めたところ,「家庭用品」において全体の購買確率が最も増加していることがわかり，その際の平均変動率は 11.9%であった．一方，他ノードの平均変動率の平均は 5.4%であった．従って，全体の購買を底上げしたい場合には「家庭用品」の購買を向上させるのが最も効率が良く，例えば「家庭用品」の購買を向上させたとき「加工食品」の購買の多い購買カテゴリは 23%増加することが分かる．

次に，より詳細な分析を行うため，小分類レベルでのネットワークにおいても同様のシミュレーション分析を行った結果を示す．ここで生起確率を変化させる小分類のノードを中分類データでの実験で平均変動率が高かった「家庭用品」に絞った．その結果，最も購買の確率が増加していた分類を 3 つ挙げると「その他家庭用品」,「台所用品」,「リビング用品」であり，その平均変動率は 1.7%，1.4%，1.3%であった．また，全ノードの平均変動率の平均は 1.1%であった．

以上のように，どの小分類の購買を増やせば顧客の購買全体を効率的に増加させることができるかを把握し，マーケティング施策の立案に結び付けることが可能となる．

5 まとめと今後の課題

本研究では，変数の階層関係を想定していなかった BN に対し，階層関係を考慮したモデルとその学習法を提案した．ノードが多数存在する場合に解釈が困難となる点について，共起関係のある 2 つの上位層ノード毎に，その共起の方向を保持しつつ下位層レベルでのネットワーク構造を作成する手法を提案し，実データへの適用によりその有用性を示した．

今後の課題として本手法を評価するために従来の手法との比較を適切な形で行うこと，観測値とするデータの最適なカテゴリ分け手法を考案することなどが挙げられる．

参考文献

[1] I Tsamardinos, LE Brown, CF Aliferis, ” The max-min hill-climbing Bayesian network structure learning algorithm, ” Machine learning, 2006.