ベイジアンネットワークのモデル構築手法の検討

鈴木　康之†　　木村　昌臣††
芝浦工業大学大学院工学研究科†　　芝浦工業大学工学部情報工学科††

1.　はじめに

不確かな事象を確率的に推論する手法としてベイジアンネットワークがある．ベイジアンネットワークは確率変数をノードで表しノード間の因果関係をアークで結んだ有向グラフであるネットワークモデルと，親ノードと子ノードの各変数がとる値の組み合わせごとに条件付き確率を割り当てる条件付き確率表に基づいて事象の予測を行う手法である．ベイジアンネットワークを利用して確率推論を行うためには事象間の因果関係を適切に反映したネットワーク構造を持つモデルを構築する必要がある．しかし，確率変数を選択するルールや変数間の因果関係を見出すルールがないため，ベイジアンネットワークを利用するユーザ自身がネットワーク構築をアドホックに行うことが多い．その場合，ユーザにより構築されたネットワークが必ずしも変数間の因果関係を反映したネットワーク構造であるとは限らない．従来の研究ではベイジアンネットワークにおける因果モデルの自動構築を提案している手法としてMWST法[1]がある．MWST法は与えられたデータからn(n-1)/2の変数間についてすべての変数間の相互情報量を求め，その中で値が大きい相互情報量が示す変数間から順にエッジを張り，n-1個のエッジが張られるまでこれを繰り返す方法であるが，確率変数間の相互情報量は変数間の関係が独立か否かのみを測る指標であるため，変数間のエッジの向きまで判別することができない．そこで，本研究では学習データから変数間の因果関係の有無だけでなくエッジの向きについての情報も得ることが可能な手法の提案を行う．

2.　因果関係抽出指標の提案

変数間の因果関係を抽出する指標として条件付きエントロピーを利用する．条件付エントロピーとは事象B（前件部）が生じているという条件下における事象A（後件部）の条件付き確率 $P(A|B)$ に関する情報量 $-\log P(A|B)$ の平均値であり，以下の式で表わされる．

$$H(A|B) = -\sum_{A,B} P(A,B)\log P(A|B) \qquad (1)$$

H(A|B) は事象Bについての知識を得た後で事象Aにまだ残っている不確定さを表す式であり，事象Bが確定したときに事象Aが一意に決まる確定的な関係に近い状況が成り立つ際に条件付きエントロピーの値は0に近づき，不確定になるほど増大する，従属性の指標として扱われる．本研究ではこの条件付きエントロピーを利用した指標を提案し，学習データから変数間の因果関係を抽出する．

Examination of the model construction method of Bayesian Networks
†Yasuyuki Suzuki　††Masaomi Kimura
†graduate school of Shibaura Institute of Technology
††Shibaura Institute of Technology

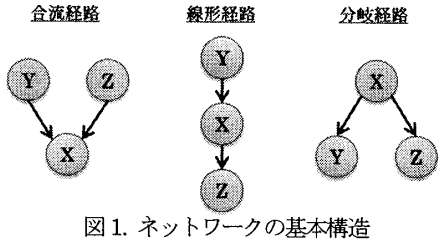

図1. ネットワークの基本構造

変数の数が多いネットワークに対してその構造を一から推定することは困難であるため，図1で示すようにネットワークの基本構造である合流経路・線形経路・分岐経路についてはじめに考える．これら3つのノードが組み合わされることでネットワークが構成されていると考え，本研究では各経路の条件付きエントロピーの値を調べる．そこでまず，条件付きエントロピーの値が最小となる式がエッジの向きを持ったネットワーク構造を表す式であるという仮説を立て，これを確かめるシミュレーションを行った．各経路の条件付きエントロピーを求めるために図1の各経路が持つ構造を基に計算して得られる1万レコードを有する学習データを人工的に生成し，ノードの状態は0と1の2値とした．各経路について生成した学習データについて表1にまとめる．

表1. 実験で用いる学習データ

	先頭ノードの出現確率	ノード間の条件付き確率
合流経路	P(Y=0)=0.50 P(Z=0)=0.50	X \ {Y,Z}: {0,0} {1,1} {0,1} {1,0} 0: 0.9 0.9 0.9 0.1 1: 0.1 0.1 0.1 0.9
線形経路	P(Y=0)=0.5	X \ Y: 0 1 0: 0.9 0.5 1: 0.1 0.5 Z \ X: 0 1 0: 0.6 0.1 1: 0.4 0.9
分岐経路	P(X=0)=0.7	Y \ X: 0 1 0: 0.8 0.5 1: 0.2 0.5 Z \ X: 0 1 0: 0.7 0.5 1: 0.3 0.5

表2. 各経路の条件付きエントロピー

H(X|YZ)	H(X|Y)	H(X|Z)
0.325（合流）	0.509（合流）	0.509（合流）
0.411（線形）	0.492（線形）	0.510（線形）
0.551（分岐）	0.567（分岐）	0.594（分岐）
H(Y|XZ)	**H(Y|X)**	**H(Y|Z)**
0.509（合流）	0.591（合流）	0.693（合流）
0.569（線形）	0.569（線形）	0.668（線形）
0.558（分岐）	0.558（分岐）	0.601（分岐）
H(Z|XY)	**H(Z|X)**	**H(Z|Y)**
0.509（合流）	0.591（合流）	0.693（合流）
0.591（線形）	0.591（線形）	0.673（線形）
0.636（分岐）	0.636（分岐）	0.652（分岐）

表1の条件を例として各経路の条件付きエントロピー求めた結果が表2である．この表より，どの経路の構造でもH(X|Y,Z)が小さくなっている．これを仮説に当てはめてみると合流経路では仮説通りであり，線形経路と分岐経路では仮説が成り立たない．よって，条件付きエントロピーの大小だけでの議論では不十分であるが，この3経路に共通しているのは変数Xを中心としてY,Zが接続することのみ分かるので，式(3)から分かることは変数の繋がり具合だけを表していると推察できる．また，一般に前件部に変数を N-1 個含む条件付きエントロピーが N 個含む条件付きエントロピーよりも値が小さくなることはないことが示せる(式(2))．(ただし，{N} はN個の確率変数を表す)

$$H(A|\{N-1\}) \geq H(A|\{N\}) \qquad (2)$$

実際，表2のように前件部に変数を2つ含む条件付きエントロピーの値が前件部に変数を1つ含む条件付きエントロピーの値よりも小さくなることが見てとれる．そこで，式(2)を考慮し後件部に変数 A を持ち前件部に変数 $N-1$ 個($N \geq 1$)を含む条件付きエントロピーと N 個含む条件付きエントロピーとの差分をとった指標を提案する．

$$H(A|\{N-1\}) - H(A|\{N\}) \qquad (3)$$

式(3)に含まれる条件付きエントロピーの前件部に余計な変数が含まれてしまうと H(A|{N})の値が大きくなり，式(3)の条件付きエントロピーの差は小さくなる．逆に余計な変数が含まれていないと式(3)の値は大きくなる．よって，式(3)の値が大きいほどネットワークの構造として H(A|{N})が示す構造が適切であると判断できる．つまり，式(3)は後件部の変数 A に N 個の変数が接続している構造が適切か，それとも $N-1$ 個の変数が接続している構造が適切かを示す指標として考えられる．これ以降式(3)を因果関係抽出指標と呼ぶ．

3. ネットワーク構築アルゴリズムの提案

本研究で提案した因果関係抽出指標を基に計算された値を用いてネットワークを構築するためのアルゴリズムを提案する．手順は次のように①～③の段階を踏む．

手順①で因果関係抽出指標に含まれる条件付きエントロピーの後件部Aに学習データ上の全変数を各々代入し，前件部Nには残りの変数を代入した式について計算を行う．次に手順②では①で計算された値の中で条件付きエントロピーの後件部の変数が等しい式同士の値を比較し，その中で最も大きい値を示す式を選択する．この選択された式が表す変数間にエッジを張る．ただし，エッジが張られた変数間の相互情報量の値が 0 となる部分の変数間にはエッジを張らない．変数間の相互情報量は前件部に変数がない条件付きエントロピーから前件部に変数が 1 つある条件付きエントロピーの差で表わされる．最後に手順③では②の操作を条件付きエントロピーの後件部の変数それぞれについて行う．

4. 実験

表1で生成した各経路の構造を基に生成した学習データに対して提案手法を適用した．表3は各経路について因果関係抽出指標を適用して計算された結果であり，表中の太枠部はアルゴリズム手順②によって選択された値である．この太枠部分の値が表す式に従って，変数間にエッジを張ると3経路ともXを中心として Y,Z が接続するネットワークが構築される．まず，線形経路と分岐経路で選択された式はいずれも相互情報量として表すことができる式であるので，方向を持ったエッジを変数間に張ることができない．一方，合流経路で選択された式には条件付きエントロピーH(X|Y,Z)を含む式が選択されている．この式が選択される理由として，変数Xを特定するためにYもしくはZが余計な変数ではなく両変数とも必要な変数であるので，XにY,Zが合流するアークを持ったネットワークが構築できる．

表3. 各経路について指標を適用した結果

H(X|Y)-H(X|Y,Z)	H(X|Z)-H(X|Y,Z)	H(X)-H(X|Y)	H(X)-H(X|Z)
0.184 (合流)	0.184 (合流)	0.102 (合流)	0.102 (合流)
0.082 (線形)	0.099 (線形)	0.120 (線形)	0.102 (線形)
0.016 (分岐)	0.042 (分岐)	0.044 (分岐)	0.018 (分岐)
H(Y|X)-H(Y|X,Z)	H(Y|Z)-H(Y|X,Z)	H(Y)-H(Y|X)	H(Y)-H(Y|Z)
0.082 (合流)	0.184 (合流)	0.102 (合流)	0.000 (合流)
0.000 (線形)	0.099 (線形)	0.120 (線形)	0.021 (線形)
0.000 (分岐)	0.042 (分岐)	0.044 (分岐)	0.002 (分岐)
H(Z|X)-H(Z|X,Y)	H(Z|Y)-H(Z|X,Y)	H(Z)-H(Z|X)	H(Z)-H(Z|Y)
0.082 (合流)	0.184 (合流)	0.102 (合流)	0.000 (合流)
0.000 (線形)	0.082 (線形)	0.102 (線形)	0.021 (線形)
0.000 (分岐)	0.016 (分岐)	0.018 (分岐)	0.002 (分岐)

5. 評価

本研究における評価として，図 2. 左で示される構造を基に計算して得られる学習データを対象に提案手法を適用し，MWST 法との比較実験を行った．

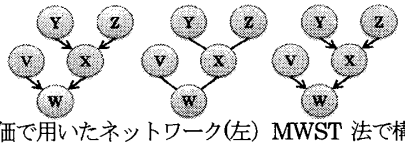

図 2. 評価で用いたネットワーク(左) MWST 法で構築されたネットワーク(中央) 提案手法で構築されたネットワーク(右)

それぞれの手法を学習データに適用した結果，MWST法では変数XにY,Zが接続し変数WにX,Vが接続した方向を持たないエッジで構成されるネットワークが求まる(図 2. 中央)．一方，提案手法を用いると条件付きエントロピーH(X|Y,Z)が含まれる式と H(W|X,V)が含まれる式が選択されることで，変数 Y,Z がXに合流し変数X,VがWに合流するネットワークが構築され元のネットワーク構造を復元できる(図 2. 右)．この結果から本研究の提案手法は MWST 法よりも合流経路を含むネットワークを復元する場合に有効な方法であることが確かめられた．

6. まとめ

本研究ではベイジアンネットワークの確率構造を推定する手法を提案した．この提案手法により，線形経路と分岐経路の判別はできないが，合流経路の判別が可能となりネットワーク構造の考え得るバリエーションの数を減らすことができると考えられる．

参考文献

[1] Chow, C.K., & Liu, C.N. Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory, IT-14, pp462-467(1968)