F-010

専門家の知識を用いるインタラクティブなベイジアンネットワーク構成手法

吉見将太† 黒川悦子† 橋本和夫†
東北大学大学院 情報科学研究科†

1 はじめに

近年，生活習慣に起因する疾患が増加しており，早期発見や予防のために，健康状態を評価する健康診断の受診やフィードバックへの関心がより高まっている．筆者らは，より効果的な支援を行うことを目的として生活習慣改善支援のためのシステム開発を行っているが，この一環としてベイジアンネットワーク (BN)[1][2] という確率モデルを用いた健康診断情報の分析の検討を進めている．BN は，循環や相関構造を持たない有向非循環グラフであり，変数間の網羅的な学習を用いてモデル構築・推論を行うため，明確な仮説がない場合でも分析が可能であるという利点を持つ．BN のモデル構築では，データの統計的偏りを情報学的基準に基づき分析してパラメータ間の依存関係を抽出するが，抽出結果の因果関係としての妥当性は必ずしも保証されない．妥当なモデル構築のためには，専門的知識も利用するインタラクティブなモデル構築手法が必要である．

本論文では，標本データのみから学習を行うフェーズと専門家の知識を導入するフェーズからなるインタラクティブなモデル構築手法について検討する．

2 専門家の知識導入の必要性

筆者らは BN を用いて生活習慣の分析 [3] を行い，歯磨き回数とメタボリックシンドロームの関連性を抽出した．一方 [4] は，口腔衛生の不良と心血管疾患のリスク上昇の因果関係を示した．このことは，筆者らが抽出した依存関係の妥当性を支持するものである．しかし，疫学分析の専門家からは，抽出されたグラフの中には，因果関係における原因と結果のリンクの向きに逆向する部分が含まれているとの指摘を受け，100% の妥当性を確認できない場合が存在した．これは，BN におけるリンクの向きは不安定なものであることを示している．与えられたデータからのリンクの向きを決定するのは不可能であり，専門家の最終判断が必要である．

依存関係を因果関係と見なすための基準は Hill の基準 [5] や Evans の条件 [6] などいくつか存在する．Hill の基準では，9 つの視点を挙げ，それらがすべて満たされた場合に因果関係として妥当であるとする

したがって，BN により因果関係導出を行うためには，抽出された依存関係を専門家の知見によって検証する必要があるためインタラクティブなモデル構築が必須となる．

3 専門家の知識導入を行うモデル構築手法

BN を用いた因果関係のモデル構築の際に，専門家の知識の導入を提案している論文がいくつかあるが，ここでは 2 つの既存手法についてまとめた．

[7] は，最初に領域データの収集を行い，これを MedKnow を用いてデータ構造を洗練する．MedKnow は，専門家に疾患とその後の経過，それらの相互作用や，周辺確率，条件付確率などを示すツールであり，専門家はデータ構造の改変を行う．最後に，Knowledge-Compiler により，先ほどの知識構造を用いた BN を構築している．この手法は，専門家の知識とデータから得られた知識を効率よく統合していると言える．しかし，2 節で述べた通り，BN のリンクの向きは不安定であり，BN が構築された後に，その構造について検証しなければ誤った因果関係を採用してしまう可能性がある．

[8] では，SFOBE(Smallest Forward-Backward Expert-Based)Model を提案している．この手法は専門家の知識によって構築されたモデルに関して再計算を行い，洗練されたモデルにする手続きであり，データから得られた依存関係に強い優先度を持たせるのが特徴である．しかし，データからの学習を重視しすぎると，抽出されるグラフは専門家の考える因果関係としての妥当性を満たさない危険性が高まる．専門的知識からではなくデータからの学習に大きく偏ってしまった場合，疫学的妥当性を満たさない可能性が高まる．また，最初に専門家の知識だけでモデルを構築しているが，専門家にも知識量の限界があるため，有用な情報の取りこぼしが十分考えられる．

以上より，BN のリンクの向きの不安定性の考慮，また，疫学的妥当性の保証と知識量の多さが必要と言える．特に後者は専門家の知識とデータの学習結果から偏りなく知識を得ることが手掛かりとなる．

A Method of An Interactive Structural Modeling using Expert Knowledge
†Shota YOSHIMI, Graduate School of Information Sciencies, Tohoku University
†Etsuko KUROKAWA, Graduate School of Information Sciencies, Tohoku University
†Kazuo HASHIMOTO, Graduate School of Information Sciencies, Tohoku University

4 An Interactive Modeling Method for Causal Analysis Incorporating Expert Knowledge

BN のリンクの不安定性を専門家の知見を用いて解消するためのインタラクティブなモデル構築手法について検討する．

Algorithm インタラクティブモデル構築手法

1: **input** グラフ G
2: $G' \leftarrow G$
3: 専門家にグラフ G' を提示
4: **while** G' で既知の因果関係以外のリンクがある **do**
5: 　G' のリンクの妥当性の検証
6: 　**if** 不安定なリンクを検出 **then**
7: 　　G' の該当箇所のリンクを修正
8: 　**else if** 不要なリンクを検出 **then**
9: 　　G' の不要なリンクを削除
10: 　**end if**
11: **end while**
12: 循環構造ができた場合は，その中で情報量の最も大きいリンクを削除する
13: G' の条件付き確率表を再計算
14: **output** グラフ G'

図 1: インタラクティブモデル構築手法の擬似コード

提案手法の擬似コードを図 1 に示した．事前準備として，分析する健康情報のデータから BN を構築しておき，1 行目ではそのグラフ G を入力とする．次に G' に G を初期値として与え，4-9 行目をグラフ G' に修正の必要がある限り繰り返し実行する．この際，専門家が提示されたグラフを読み取り，既知の因果関係以外のリンクがあるかどうかの判断を下す．既知の因果関係以外のリンクがある場合，5 行目のリンクの妥当性の検証を行う．リンクの妥当性の検証は，2 で挙げた Hill の基準や Evans の条件に基づいた検証を行う．6，7 行目は，G' において専門家が不安定だと判断したリンクがあれば，G' の他の部分はその構造を保持したまま不安定箇所のみを修正する．8，9 行目は，妥当性がない不要なリンクを検出した場合はそれを削除する．この操作も，G' の他の部分はその構造を保持したまま行う．これらの操作を既知の因果関係以外のリンクがなくなるまで繰り返した後，循環構造の有無を確認する．循環構造がある場合，その循環構造の中で情報量が最も大きく，当てはまりの悪いリンクを削除する．次に，G' の条件付き確率表を再計算する．最終的に，入力したグラフ G は，グラフ全体において因果関係が示された因果構造グラフ G' として出力される．

以上が提案手法の概要である．既存手法と比較すると，提案手法は専門家の知識の導入を重視しており，グラフにおけるリンクの不安定性の解消に重点を置いている．

5 人間ドックのデータからのモデル構築

本章では，4 で提案したモデル構築手法について，専門家の知識を導入する際の手順確認，および提案手法の有用性の確認のため，人間ドックのデータを用いた評価実験を行う．ただし，本論文ではこの実験を予備実験と位置付ける．専門家の知識を導入する際の手順確認，提案手法の有用性の確認のためである．

5.1 分析対象

分析対象は，2008 年に実施された宮城県内での人間ドックのデータ 13,979 件である．対象人数は約 14,000 人で十分なサンプル数であり，広範囲な受診者層であることから，特定の項目で偏りが生じている可能性は低い．なお，データの取り扱いは，個人情報をすべて削除した状態で行った．

5.2 標本データの学習

ベイジアンネットワークのモデル構築には，Visual Mining Studio というソフトウェアを用いる．同ソフトウェアは Bayesian Network Module をアドオンすることにより，ベイジアンネットワークを用いたモデル構築をすることが可能となる．

計算機環境は以下の通りである．

CPU ：Intel Xeon(R) X5460 @3.16GHz

メモリ ：8GB

OS ：Windows XP(64bit オペレーティングシステム)

5.3 専門家の知識の導入

本実験では，専門家の知識の代わりとして，医学的研究結果を用いることとする．医学的研究結果は，疫学的に妥当性のあるものであり，専門家の知識の代わりになり得るものである．

5.4 実験結果

図 2 に標本データの学習により構築された BN を示した．ノード内の数字が表 1 の健診・問診項目に対応している．また，医学的研究結果を用いて修正した BN を図 3 に示す．

図 2 に対して，医学的研究結果を適用したが，これは [9][10] を参考にしたものである．医学的研究は高血圧や高血糖などの疾患と生活習慣との関わりを示しているものがほとんどであり，生活習慣同士の関わりを示すものではない．したがって，本実験でも疾患と生活習慣とのリンクが生成された部分のみに着目した．

図 2 において，疾患と生活習慣との間にリンクが生成された部分は，$2 \rightarrow 8$，$3 \rightarrow 8$，$14 \rightarrow 1$，$14 \rightarrow 2$ であった．ここで，$14 \rightarrow 1$，$14 \rightarrow 2$ に関しては因果関係が保証されてい

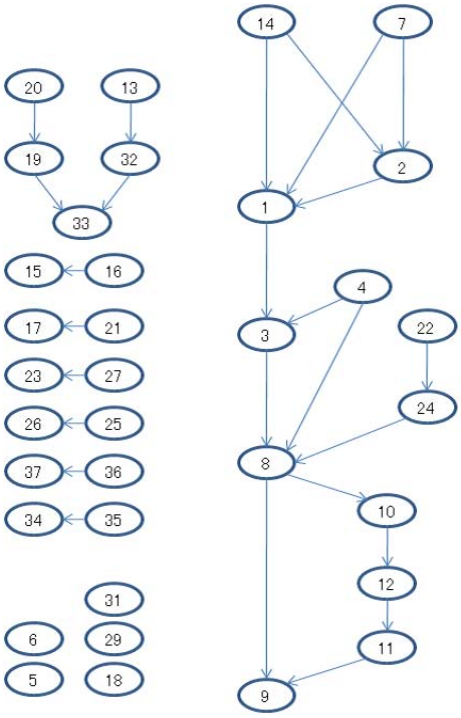

図 2: 標本データの学習で構築された BN

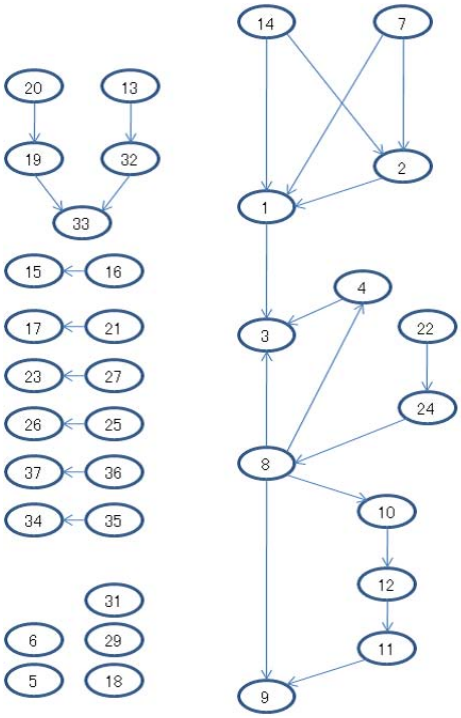

図 3: 医学的研究結果を用いて修正した BN

表 1: 健診・問診項目

	健診結果
1	BMI
2	腹囲
3	収縮期血圧 (最高血圧)
4	拡張期血圧 (最低血圧)
5	中性脂肪
6	空腹時血糖
7	HDL コレステロール
	質問項目
8	飲酒の頻度はどのくらいか
9	喫煙習慣はあるか
10	歯科検診を定期的に受けているか
11	1 日に 2 回以上歯磨きをするか
12	歯間部清掃用具を使用しているか
13	1 日におよそ何分くらい歩くか
14	18-20 歳頃の体重に比べて増減はあるか
15	食事の速さはどうか
16	お腹いっぱい食べることがあるか
17	食事時刻は規則的か
18	朝食は週何日くらい摂るか
19	昼食が外食となる日は週何回あるか
20	夕食が外食となる日は週何回あるか
21	夕食を食べてから寝るまで何時間あるか
22	夕食後何か食べることは週何回あるか
23	栄養のバランスを考えて食事をしているか
24	菓子類，糖分の入った飲料を摂るか
25	脂肪分の多い食事を摂るか
26	食事の塩味はどうか
27	野菜の量はどうか
28	栄養成分の表示を参考にするか
29	カルシウムに富む食品を食べるか
30	運動不足だと思うか
31	仕事以外の時間に汗をかくような運動をしているか
32	日常における身体活動はどうか
33	職種は何か
34	休養は充分であると思うか
35	睡眠は充分であると思うか
36	朝目覚めたときに爽快感を感じるか
37	ストレスがたまっていると感じることがあるか

るので，リンクの修正は行わない．しかし，2 → 8，3 → 8 に関しては，[9][10] の結果より，リンクの不安定性が表れており，逆向するリンクの方が妥当性があると考えられる．肥満に関係するのは，中性脂肪や血糖よりも自己測定可能な血圧という事で，その肥満の大元は体重増加であり，血圧の生活習慣は飲酒が大きく左右する．したがって，8 → 2，8 → 3 となるようにリンクの向きを修正した．最終的に図 3 にある BN が導かれた．またこの結果から，飲酒と夜遅い食事や菓子類との関係性，喫煙と歯磨きなどの口腔衛生が関係していることがわかった．

5.5 実験結果の検証

再構成された BN が図 3 に示されたが，これは情報量規準にしたがった BN の定義からも大きく逸脱しておらず，かつ疫学的な妥当性も保証されていると考えられる．はじめに標本データの学習を行ったが，リンク生成するための情報量の閾値を比較的高めに設定したため，リンクが生成されにくい状況であった．したがって，構築された BN において疾患と生活習慣との間に生成されるリンクが少なかった．また，疫学的に妥当性のないリンクも少なく，修正の必要があまりなかったために，情報量基準から大きく外れることはなく，BN 全体で見ても疫学的妥当性があると言える．提案手法の手順確認についても，今回は円滑に進めることができ，問題はなかったと言える．

しかし，生成されるリンクが少ないということは，新たな知識の発見や，複雑に交絡した事象を明確に表現することに繋がりにくいことがある．ただし，リンクを増やしすぎると専門家の知識を導入するフェーズで，大きく時間を費やし，専門家の判断ミスが生じる可能性も出てくる．どの程度リンクを生成するかが，今後の大きな課題になると考えられる．

以上より，いくつかの課題はあるが，本実験において提案手法の有用性を確認することができた．

6 まとめ

本論文では，標本のデータのみから学習を行うフェーズと専門家の知識を導入するフェーズのインタラクティブなモデル構築手法について提案した．また，実際に専門家の知識を導入する前の予備実験を行った．

今後は，5.5 で述べた課題に取り組み，それに付随して専門家の知識量や正確さ，リンクの妥当性を判断する所要時間などを予備実験において再現する必要がある．専門家の知識獲得手段を定義し，アルゴリズムに盛り込まなければならない．その上で実際に専門家の知識を導入したモデル構築の実験に移りたいと考えている．

謝辞

本研究の一部は，文部科学省の平成 19 年度知的クラスター創成事業 (第 II 期) の助成を受けて実施したものである．

同事業の研究統括である後藤順一教授，東北大学大学院医工学研究科永富良一教授には，本論文をまとめる際に様々なアドバイスを頂いた．ここに感謝する．

参考文献

[1] David Heckerman, “A Tutorial on Learning With Bayesian Networks”, *Technical Report*, MSR-TR-95-06 (1995).

[2] Judea Pearl, “統計的因果推論”, 共立出版株式会社 (2009).

[3] 吉見将太, 黒川悦子, 橋本和夫, “ベイジアンネットワークを用いた生活習慣の分析”, 第 9 回情報科学技術フォーラム, G-029 (2010).

[4] Cesar de Oliveira, Richard Watt and Mark Hamer, “Toothbrushing, inflammantion, and risk of cardiovascular disease: results from Scottish Health Survey”, *BMJ*, 340:c2451 (2010).

[5] Austin Bradford Hill, “The Environment and Disease: Association or Causation?”, *Proceedings of the Royal Society of Medicine*, 58, 295-300 (1965).

[6] ALFRED S. EVANS, “Causation and disease: The Henle-Koch postulates”, *Yale J Biol Med*, 49: 175-194 (1976).

[7] J. Horn, T. Birkholzer, O. Hogl, M. Pellegrino, R. Lupas Scheiterer, K.-U. Schmidt, and V. Tresp, “Knowledge Acquisition and Automated Generation of Bayesian Networks for a Medical Dialogue and Advisory System”, *Artificial Intelligence in Medicine*, Springer, 199-202, (2001).

[8] Ruxandra Lupas Scheiterer, Dragan Obradovic and Volder Tresp, “Tailored-to-Fit Bayesian Network Modeling of Expert Diagnostic Knowledge”, *Journal of VLSI Signal Processing*, 49, 301-316 (2007).

[9] 若林一郎, “定期健康診断の結果からみた山形県内の産業従事者の血中脂質異常について”, 秋田県公衆衛生学雑誌 第 2 巻 第 1 号 7-11 (2005).

[10] 横山裕一, “過剰飲酒者におけるメタボリックシンドローム患者の増加の背景”, 日本アルコール・薬物医学会雑誌 42(4), 354-355 (2007).