G-020

ベイジアンネットワークを用いた生活習慣分析

Lifestyle Analysis Using Bayesian Network

吉見将太 † 黒川悦子 ‡ 橋本和夫 †
Shota YOSHIMI Etsuko KUROKAWA Kazuo HASHIMOTO

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，生活習慣の改善に対する関心が高まっている．生活習慣病など様々な疾患の早期発見や予防のために，健康状態を評価する健康診断がある．健康診断では，生活習慣を把握するために様々な問診が行われるが，健診結果と生活習慣との複雑な関係は十分に把握されているとは言えない．理由としては，まず，その関係について詳細なフィードバックが行われていない，もしくは少ないことが挙げられる．また，フィードバックがあったとしても，専門家でなければ知見の導入が難しいということも言える．

問診結果と健診結果の関係については，項目間の依存関係をグラフィカル表示することにより，視覚的なモデル把握が容易となる．

ベイジアンネットワーク [1][2][3][4] は有向非循環グラフであり，循環や相関構造を持たない．また，変数間の網羅的な学習を行いモデル構築・推論を行い，かつ，明確な仮説がないところから分析が可能である．そこで，本論文ではベイジアンネットワークという確率モデルを用いて生活習慣に関する問診結果の分析を行う．

2 既存手法

2.1 統計的検討

あるデータに対し，疫学における統計的な検討を与える手法は多岐にわたる．参考文献 [5] では，統計学を記述統計学と分析統計学に大分している．記述統計学とは，収集したデータを解釈しやすいように提示するものであり，分析統計学は推定や検定などの統計学的推論を表している．

健康診断などの結果について見解を与える論文については，記述統計学，分析統計学ともによく用いられる．ただし，記述統計学は標本となる観察対象集団に対してのみの状況が得られるだけである．一方，分析統計学は標本を通じて標的集団である母集団の状況を推論する．したがって，大きな母集団に適用できるモデルを導入しようとする場合，母集団全体からデータ収集を行うことは非常に困難であるため，統計学的推論を使うことになる．多くの論文では，ある狭い範囲の標本から得られたデータから統計学的推論を行い，母集団の本来の状況を推し量ろうとしている．

統計学的推論には，オッズ比の算出や多変量解析などの統計学的推定と，t 検定やカイ 2 乗検定などの統計学的検定があるわけだが，専門的な知識がなければ知見の導入が難しい場合が多い．専門家向けの分析結果に近いと言える．また，メタボリックシンドロームと生活習慣の関係といったように，複雑に交絡した項目間の関係を明確に記述する手法も少ない．そこで，専門的な知識を持っていなくても，ある程度の知見を導入でき，かつ複雑に交絡した項目間の関係を表す方法として，モデルをグラフィカル表示することが考えられる．

モデルをグラフィカル表示できる手法としては，共分散構造分析やパス解析 [6] などがある．しかし，これらの手法では因果関係の循環などがモデルに入ってしまう場合がある．この変数間の循環構造が存在すると，モデルの解釈が難しくなる．また，共分散構造分析は明確な仮説がなければモデルを構築できないという欠点がある．健診・問診結果などから明確な仮説をたてることは容易ではない．

まとめると，モデルのグラフィカル表示など知見を導入しやすく，かつ，健診・問診結果のように項目が複雑に交絡したデータに適した手法はこれまでに用いられていなかったと言える．

2.2 分析対象となる集団

疫学的立場から分析を行う論文は多々ある．しかしながら，大きな標本からのデータ収集の難しさなどがあり，学校内，企業内などの比較的狭い範囲の標本からのみデータ収集をする場合が少なくない．データ収集を行う標本があまりにも小さい場合，生活習慣や環境が似ているため，結果に偏りが生じてしまい，統計学的推論を行っても他の集団へも適用可能な汎用性のあるモデル構築ができない場合がある．

2.3 データの取り扱いに関して

情報学的立場から

2.1 で述べたように，疫学において統計的な検討手法は多い．したがって，手法によっては，詳細なデータの操作が

†東北大学大学院 情報科学研究科, Graduate School of Information Sciencies, Tohoku University
‡東北大学大学院 医工学研究科, Graduate School of Biomedical Engineering, Tohoku University

必要となる．例えば，項目 a を独立変数，項目 b,c,d を従属変数として重回帰分析を行うように指示するような操作がそれにあたる．データの中で，分析したい箇所が一部分だけであれば問題はないが，データ全体から有用な情報を抽出しなければならない場合，逐一条件設定などを行えばとりこぼしが生じる．計算量が増大してしまう，分析ツールに実装されていないなどの理由から網羅的な学習を行っていない場合もある．

疫学的立場から

データの形式には，2 値データ，数段階のリッカート尺度によるデータ，カテゴリデータ，連続値などさまざまなものがある．問診などでは，リッカート尺度が用いられることがよくあるのだが，注意しなければならない点がある．順序尺度として用いる場合，必ずしもその順序で変数の説明に用いることが出来ないことである．例えば，「他人と比較して酒を飲むほうである」という問診に，「よくあてはまる」から「まったくあてはまらない」までの 5 段階リッカート尺度で応答するとする．もし，少量の飲酒は，まったく酒を飲まない場合よりも体に良いのであれば，この順序尺度はこのままで使うことはできない．

このように，順序尺度があるような問診結果は扱い方に注意する必要がある．この点を考慮していない論文も見受けられる．

3 分析手法

ベイジアンネットワークは以下の特徴を持つ．

- 複雑に交絡した変数を扱うことができる
- 網羅的な学習により有用な情報のとりこぼしがない
- 順序尺度の概念を考慮する必要がない
- モデルがグラフィカル表示される

順序尺度の概念を考慮する必要がないというのは，ベイジアンネットワークではデータをカテゴリ型のデータとして扱うためである．

本論文では，健診・問診結果をベイジアンネットワークを用いて分析する．

具体的な分析手法は以下に示す．

3.1 分析ツールと計算機環境

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．このソフトウェアは種々の統計的な分析ができる汎用ツールである．アドオンとして Bayesian Network Module を導入することにより，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．

計算機環境としては，データのサイズが大きくなるほど計算量が増大してしまうので，ある程度スペックの高いマシンが要求される．本論文では以下の環境で分析を行った．

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)

3.2 分析手順

分析は以下の 1，2 の順で行われる．2 つのフェーズを採用することにより，計算量減を図る．

1. 独立性探索フェーズ
   (a) ノード間の依存関係による向き付けを探索する
   (b) 情報量規準に基づく条件付独立性とグラフ理論の適用を行う
2. リンク探索フェーズ
   (a) 向き付けされたノード間のリンクの有無を探索する
   (b) 独立性探索フェーズで決まらないノード間の連結を，情報量規準に基づくスコアによりネットワーク構造を評価し，Greedy 探索によりノード間のリンク構造を探索する

3.3 分析で用いる情報量規準関数

3.2 で用いた独立性探索フェーズとリンク探索フェーズでは，独立性の判定とリンクの有無の判定をするために，以下で定義される情報量規準関数 [7][8][9][10] を用いる．本論文ではベイズ情報量規準を採用するが，様々な分野でよく利用される赤池情報量規準 [11] を比較のためにともに示す．

ベイズ情報量規準 (BIC)

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n) \quad (1)$$

L ：尤度関数

k ：パラメータ数 (自由度，独立変数の数)

n ：標本の大きさ

赤池情報量規準 (AIC)

$$AIC = -2 \cdot \ln(L) + 2 \cdot k \quad (2)$$

L ：尤度関数

k ：パラメータ数 (自由度，独立変数の数)

この情報量規準は，モデルの良さ (正確さと複雑さのバランス) を評価するための尺度である．ベイズ情報量規準，赤池情報量規準ともに，情報量が小さいほどよいモデルである．

また，情報量規準関数によってはモデル構築の傾向が変わる．本論文で分析した際に用いた情報量規準関数はベイズ情報量規準であり，これは赤池情報量規準などと比べ，母数が少なめのモデルを選択する傾向がある．このため，情報量規準関数の選択と，判定のための情報量の閾値の設定は，任意に行う必要があるといえる．

表 1: 質問項目群

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

3.4 目的と分析対象

本論文ではメタボリックシンドロームと生活習慣の関係を分析することを目的とする．健診結果からメタボリックシンドローム判定を対象に加えた理由としては，好ましくない生活習慣が起因して発生する疾患としての指標となりうるからである．

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ている．

本論文の分析に利用した項目群を表 1 に示す．健診結果はメタボリックシンドローム判定，基準該当・予備群・非該当の 3 区分，質問項目は 32 項目を用い，質問項目 2,3,4 は回答の選択肢が 2 つ，質問項目 5 から 8 は選択肢が 3 つ，それ以外の質問項目は 4 つの選択肢が与えられる．表 1 では選択肢は割愛してある．

4 分析結果

構築されたモデルを図 1 に示す．ノード内の数字は表 1 に示した質問項目群の番号に対応している．

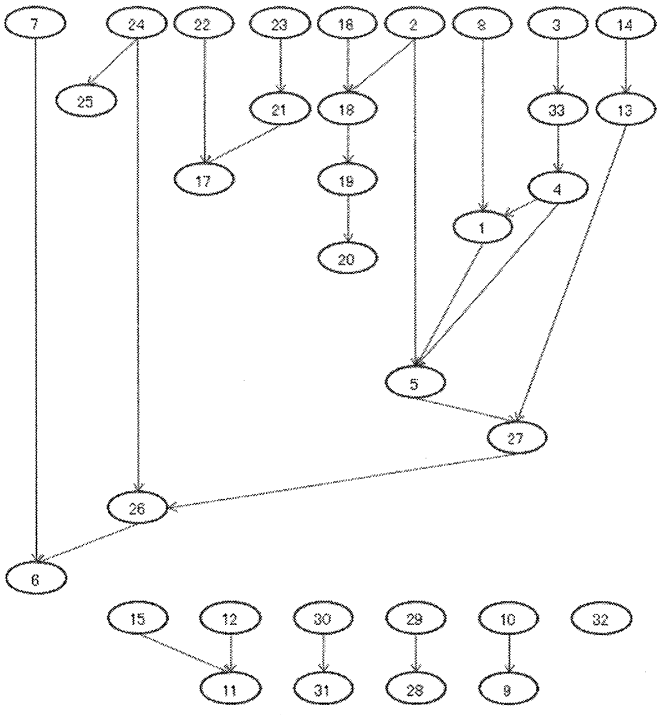

図 1: 構築されたモデル

また，構築されたモデルから得られた結果を以下で述べる．

- メタボリックシンドローム判定は，項目 3, 4, 8, 33 の子ノードとなっており，「1 日 2 回以上の歯磨きをするかどうか」，「18-20 歳頃の体重に比べて増減があるか」という質問への応答が，メタボリックシンドローム判定へ影響している
- 問診項目同士の依存関係が存在している
- 項目 32 は唯一，他のノードとリンクがなく，独立な項目である

ここで，このモデルに疫学的な見解を入れる．イギリス医師会雑誌 (BMJ) で 2010 年 5 月に発表された論文 [12] では，歯磨き (口腔衛生の不良) が心血管疾患のリスクの上昇に関連すると述べている．本論文で構築したモデルでは，歯磨き回数がメタボリックシンドローム判定に寄与していることがわかる．メタボリックシンドロームは心疾患と強い依存関係にあることから，このモデルは疫学的にも妥当性のあるものだと推測する．この点で，ベイジアンネットワークによるモデル構築は，有用な情報をとりこぼすことなく抽出できているといえる．

5 まとめ

ベイジアンネットワークを用い，メタボリックシンドローム判定と生活習慣に関する問診結果のモデル構築を行った．その結果として，次の 3 つを実現することができた．

- メタボリックシンドロームと生活習慣に関する問診結果の複雑な関係のモデル化
- 疫学的にも妥当性のある依存関係の抽出
- モデルのグラフィカル表示

これにより，ベイジアンネットワークによる分析が有効であることが確認できた．

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．

同事業後藤順一研統括，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] Yoichi MOTOMURA, “Bayesian Networks”, *TECHNICAL REPORT OF IEICE.*

[2] Hiroki SUYARI, “Introduction to Bayesian Network”, *MEDICAL IMAGING THCHNOLOGY Vol.21 No.4*, September 2003.

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

[4] C.M. ビショップ, “パターン認識と機械学習　下”, シュプリンガー・ジャパン株式会社, 2008.

[5] 中村好一, “基礎から学ぶ楽しい疫学　第 2 版”, 医学書院, 2006.

[6] 狩野裕, “構造方程式モデリングは，因子分析，分散分析，パス解析のすべてにとって代わるのか?”, 行動計量学, 29, 138-159, 2002.

[7] Moninder Singh and Marco Valtorta, “Construction of Bayesian Network Structure from Data : a Brief Survey and an Efficient Algorithm”, *International Journal of Approximate Reasoning*, 1995.

[8] Moninder Singh and Marco Valtorta, “An Algorithm for the Construction of Bayesian Network structure from data”, *Proceedings of the 9th Conference on Uncertainty in Artificial Interigence*, 1993.

[9] Gregory F.Cooper and Edward Herskovits, “A Bayesian Method for the Introduction of Probabilistic Networks from Data”, *Machine Learning*, 1992.

[10] Herman J.Bierens, “Information Criteria and Model Selection”, (Pennsylvania State University) 2006.

[11] 赤池弘次, “赤池情報量規準 AIC”, 共立出版株式会社, 2007.

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