因果関係の可視化を考慮したベイジアンネットワークのベイズ最適な予測法

1715038 清水　豪士

富山県立大学　情報基盤工学講座

October 18, 2019

はじめに

背景

近年、情報技術の発展に伴い、データマイニングやパターン認識の技術が注目を集めている。
現在、因果モデルと確立予測モデル、それら両方の特徴を持つ手法であるベイジアンネットワーク（以下、BN と表記する）の研究が盛んである。

目的

- ベイズ最適な予測精度を持ちつつ、混合モデルの確立構造の解釈容易性を伴った、BN の混合予測モデルの構成法を提案する。
- 新たに相互情報量を用いて、ノード間の因果関係の強さを定量的に評価する方法を提案する。

はじめに

従来研究

- 真のモデルに近づけることを目的にモデルの混合を行っているモデルクラスを探索的に与え、そのうえで混合を取る方法を与えている
- 従来研究で提案されたモデルは、各モデルの重要度を考慮せずに混合しているため、予測精度の低いモデルの影響を受け、逆に予測精度を下げてしまう恐れがある

ベイジアンネットワークとは

ベイジアンネットワークとは、複数の確率変数間の定性的な依存関係をグラフ構造によって表し、個々の確率変数間の定量的な関係を条件付き確率で表した確率モデル

下図において X_1 は X_2 X_3 の親ノードという
また、X_2 、X_3 は X_1 の子ノードという

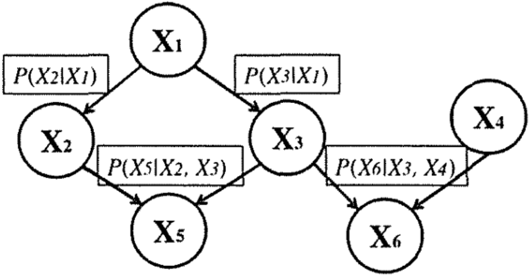

図 1　ベイジアンネットワーク例

BN モデルの特徴

BN モデルの特徴

1. 確率変数間の因果関係をグラフィカルなネットワークで表現可能
2. ある観測値データが与えられたもとで、その他の変数の条件付き確率を推論する現実的な計算アルゴリズムが与えられている
3. 確率モデルとしての整合性を持った推論が可能

問題設定

学習データの総数を N 、確率変数の総数が n の観測値からなるデータ D =(x^1 , ⋯ ,x^n), x^i = (x_1^i ,⋯, x_n^i) から、BN モデルを統計的に学習し、モデル上の一部の確率変数の観測値を入力したもとで状態が未知の確率変数に対して事後分布を出力する確率推論を行い予測値を得る

X_q : 親ノード　X_p : 子ノード とする
親ノードが複数の時、子ノード X_q の親ノードの集合を $\mathcal{P}_a(X_p)$
全変数を合わせた n 次元確率変数 X=(X_1 ,⋯, X_n) の同時確率分布を (1) 式で表現する

$$\boldsymbol{P}(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i|\mathcal{P}a(X_i)) \qquad (1)$$

最尤推定法

1 つのモデル T_l (l = 1, ···, m) が与えられたとき、 N_{ijl} を学習データのうち X_i の親ノード集合が j 番目のパターンであるもとで、X_i が k 番目の値をとるデータ数と定義する。

条件付き確率の最尤推定量 $\hat{\theta}_{ijk}^{T_l}$ は (2) 式で表現する

$$\hat{\theta}_{ijk}^{T_l} = \frac{N_{ijk}}{\sum_k N_{ijk}} \tag{2}$$

確率推論

確率推論とは、確率変数の値が観測されたもとで、構築したモデル構造を利用し、目的変数の事後確率を求めること。

$$P(X_i|e^+,e^-) = \frac{P(X_i|e^+)P(e^-|X_i)}{P(e^-|e^+)} \tag{3}$$

e^+：X_i の親ノードの観測情報　　e^-：子ノードの観測情報

提案手法

本研究では、何らかの方法に基づいてノードの先行順序関係が与えられるものとする

ノードの先行順序関係の与える方法

1. 分析者が対象問題の技術的知見に基づいて因果関係を決定する
2. 全ノード間でリンクを仮定し、そのリンクの方向の組み合わせについてモデル選択により決定する
3. G^2 検定によりノード間の結合の有無を検定した後、全てのリンクの方向の組み合わせについてモデル選択により決定する

本提案手法はこのようなノード間の先行順序関係のみが与えられたもとで、最も複雑なモデルの部分集合となるモデル全てで混合を取る方法である

依存度合の可視化

提案した BN の混合モデルは最も複雑であるため、ノード間の因果関係を視覚的に解釈することができない

↓

相互情報量を用いて、ノード間の因果関係の強さを定量的に評価する

相互情報量とは、2 変数の因果関係の強さを計量化した尺度

$$I(X_p; X_q) = \sum_{x_p, x_q} P(x_p, x_q) \log \frac{P(x_p, x_q)}{P(x_p)P(x_q)} \quad (13)$$

予測対象

本研究では、日経平均株価のゴールデンクロス (GC) 及びデッドクロス (DC) と呼ばれる株の売買指標を予測対象とする

予測対象

- ゴールデンクロス (GC)
  短期移動平均線が長期移動平均線を上回った点
- デッドクロス (DC)
  短期移動平均線が長期移動平均線を下回った点

実験条件

本実験では、ダウ工業平均株価（米）、Financial Times Stock Exchange100(FTSE)（英）、日経平均株価（日） の 3 指標の週データを用いる

これらのデータは時差があるため、「日経> FTSE >ダウ」という先行順序関係をもとに、各ノードの親ノード候補集合を決定する

実験条件

実験に用いるノード

1. 前週と比較して日経株価が上昇・下降の 2 値
2. 4 週間前と比較して日経株価が上昇・下降の 2 値
3. 長短期の移動平均線のどちらが大きいかの 2 値
4. 長短期の移動平均線の差 500 以上・以下の 2 値
5. 前週と比較してダウ株価が上昇・下降の 2 値
6. 前週と比較して FTSE 株価が上昇・下降の 2 値
7. 1 の翌週の値
8. 2 の翌週の値
9. 3 の翌週の値

以上の 9 つのノードを用いて、GC 発生、DC 発生、発生しない の 3 値を持つノードを予測する

モデルの学習期間は 5 年、予測期間は 1 年とし、これを期間を 1 年ずつずらして 3 セットの実験を行う (2007〜2009 年)

実験条件

予測の際は、現在の値を表すノードを観測情報として (3) 式を用いて確率推論を行い、"GC 発生"、"DC 発生"、"発生しない" の 3 つの事象の事後確率を求める

多くの平常時では GC,DC のどちらも発生しないため、"発生しない" の事後確率が高くなってしまう
GC 発生, または DC 発生の確率が、ある閾値を超えた時に発生すると判定する

本稿では、GC 発生、または DC 発生の確率が 0.4 を上回った場合を発生する と判定する

実験条件

実験は以下の 3 手法で比較を行う

- BIC を用いた提案手法 (提案手法)
- Ammar らの混合方法（従来研究） (Ammar)
- BIC 基準でモデルを 1 つ選択する方法 (BIC)

各手法で 3 年分のデータに対して予測を行い、以下の指標により評価する

$$\text{正解率} = \frac{\text{正解した回数}}{\text{GC(DC) と予測した回数}} \tag{14}$$

この場合の「正解」とは、GC,DC が発生すると予測してから 7 週間以内に発生した場合のこと

結果

2007 年を予測年とする際に用いた提案法による混合モデル

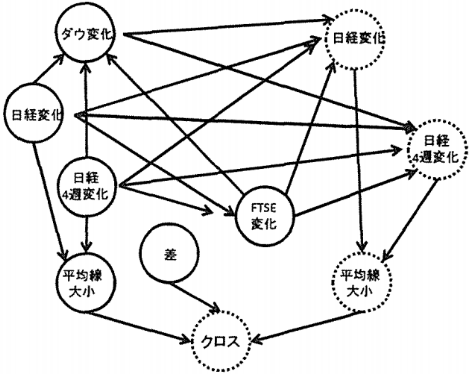

図 3　提案法による混合モデル（2007 年予測）

実線 ： 現在のノード　　点線 ： 翌週 もしくは 目的変数のノード

結果

2007 年を予測年とする際に用いた BIC により選択されたモデル

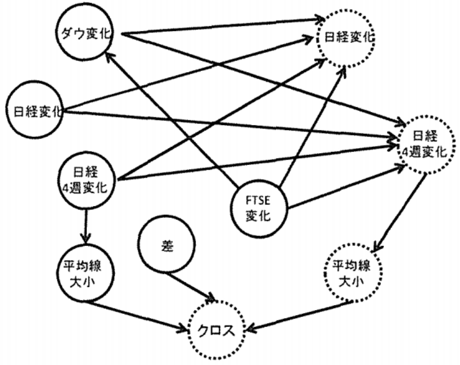

図 4　BIC により選択されたモデル（2007 年予測）

実線 ： 現在のノード　　点線 ： 翌週 もしくは 目的変数のノード

結果

2007 年を予測年とする際に用いた Ammar らの手法による混合モデル

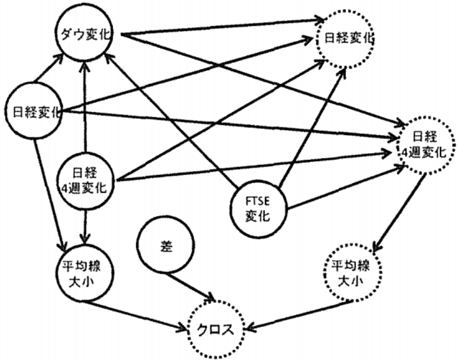

図 5　Ammar らによる混合モデル（2007 年予測）

実線 ： 現在のノード　　点線 ： 翌週 もしくは 目的変数のノード

結果

2007～2009 年を予測年とした場合の GC と DC の正解率である

表 2 GC の予測実験結果

予測年	BIC	Ammar	提案手法
2007	80.0%	80.0%	87.5%
2008	40.0%	40.0%	55.6%
2009	61.1%	57.9%	65.0%
計	60.53%	58.97%	67.57%

表 3 DC の予測実験結果

予測年	BIC	Ammar	提案手法
2007	83.3%	90.9%	87.5%
2008	100.0%	100.0%	100.0%
2009	80.0%	73.3%	83.3%
計	85.00%	82.76%	87.10%

結果

BIC による漸近事後確率に基づいてモデルを混合した方が BIC でモデル選択を行う方法、 Ammar らの重要度を考慮しない方法と比較して優れた予測精度が得られる

閾値について

表 4　GC の予測実験結果

閾値	BIC	Ammar	提案手法
0.3	58.97%	57.50%	65.79%
0.4	60.53%	58.97%	67.57%
0.5	84.21%	80.00%	100.0%

表 5　DC の予測実験結果

閾値	BIC	Ammar	提案手法
0.3	58.70%	60.87%	62.75%
0.4	85.00%	82.76%	87.10%
0.5	(予測なし)	(予測なし)	(予測なし)

- 閾値が大きい場合
  精度が高まるが、 DC・GC の発生の予測がされ難い
- 閾値が小さい場合
  DC・GC の発生の予測が頻繁にされるようになり、精度は低下する

結合強度の可視化 1

提案法の混合 BN モデルにおいてノード間の相互情報量を付与する

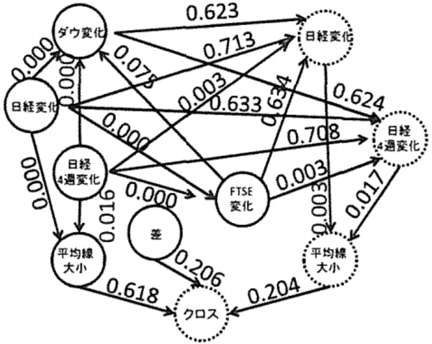

図 6　相互情報量を付与した提案法の混合モデル

結合強度の可視化 2

図 6 の関係性をより強調して可視化したもの

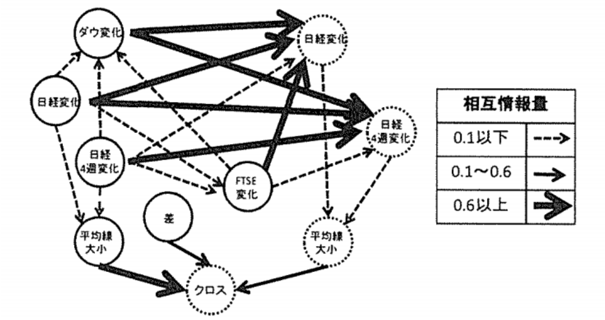

図 7　ノード間の関連性を可視化した混合モデル

図 6 のものと比べ、解釈容易性が向上している

考察

Ammar らの手法は、正解率の低いモデルの影響を強く受けたと考えられる

提案した混合方法は事後確率でモデルを混合しているため、BIC でモデルを 1 つ選択したものよりも優れた正解率が得られたことから有効性が示せた

図 7 のように相互情報量を用いてグラフの可視化を行うことで、ベイズ最適な予測精度を保ちつつ、ユーザが任意で定める閾値以上のリンクを注目し、混合モデルの因果関係の構造を把握することが容易になった

まとめと今後の課題

まとめ

- 提案法により、BN モデルを用いてベイズ最適な予測精度の高い予測が構成できた
- 同時に、因果関係の解釈容易性という本来備えているモデルの長所を有した方法を与えることができた

今後の課題

- BIC による事後確率の漸近近似値ではなく厳密解での混合方法の導出、他のデータを用いた評価実験による検証