アンサンブルを用いたベイジアンネットワークの構造学習に関する研究

1X11C118-5　矢野涼太
指導教員　後藤正幸

1　研究背景と目的

近年，情報技術の発達に伴って大量のデータが生成・蓄積されており，そのデータをどう活用するかが重要な課題となっている．そこで，そのようなデータの要素間の因果関係の可視化や未観測の要素における事象の発生確率の予測が行えるベイジアンネットワーク (以下，BN) が注目されている [1]．BN は要素をノードで表し，因果関係にあるノード同士を非循環有向リンクで結ぶことで全体をネットワーク構造で表現し，一部のノードの事象について観測値が得られたもとで，未観測のノードの事象の値を予測することができるモデルである．以下，因果関係にあると推定できるすべてのノード間に有向リンクを作成した状態を，リンク構造と呼ぶ．BN の学習における推定精度向上のための方法の 1 つとして，ノード間の因果関係を適切に表現したリンク構造を持つ有向グラフを作る方法が考えられる．この考え方に従い，複数のモデルを混合するアンサンブルを用いることで，真の構造を推定しようとした手法が Ammar らによって提案されている [2]．Ammar らの手法ではまず，サブデータセットを用いて無向リンクの集合である無向グラフを複数作成し，それらをアンサンブルした混合無向グラフを作る．さらに，この混合無向グラフに含まれる無向リンクを有向リンク候補として，全学習データを用いてその中から有向リンクを探索的に選択し，リンク構造を推定する．しかし，最も重要なリンク構造の推定は，混合無向グラフ上で全学習データによって行われるため，アンサンブルは無向グラフの推定にのみ用いられる．そのため，学習データに過適合し，ノード間の方向性までを考慮した因果関係を適切に推定できないという問題が発生する．

そこで本研究では，各サブデータセットのもとで BN のリンク構造を推定し，これらのリンク構造に対してアンサンブルを行うことで，ノード間の因果関係を適切に推定したリンク構造を持つ有向グラフを作成する方法を提案する．本提案は，ノード間の因果関係を適切に推定することで，確率推論における予測精度を向上させることも期待できる．また，提案手法の有効性をベンチマークデータに対する実験により検証する．

2　ベイジアンネットワーク

2.1　ベイジアンネットワークの概要

BN は，構造学習と確率推論の 2 つのステップから構成されている．まず，構造学習では，分析対象の各要素 (確率変数) をノードで表現し，ノード間の因果関係を有向リンクによって表現し，リンク構造を作成する．有向リンクとは，因果関係にある 2 つのノード間に作られる，向きを持った関係性を表し，通常は付与される条件付き確率の向きとなる．その後，確率推論によって観測値が与えられたノードの事象の値から，未観測のノードの事象について発生確率の予測を行う．

2.2　構造学習

BN における代表的な構造学習の手法として，Greedy Hill-Climbing がある．Greedy Hill-Climbing では有向リンクが 1 つも存在しない状態からスタートし，下記の中から Bayesian Information Criterion(BIC) スコアが最も高くなる操作を繰り返し行う手法である．

- 任意の有向リンクを 1 つ追加
- 任意の有向リンクを 1 つ削除
- 任意の有向リンクを 1 つ反転

BIC スコアとは，データに対するモデルの適合度を表すモデル選択のためのスコアである．

また，Greedy Hill-Climbing は単体で用いると高次元データの場合に計算量が膨大になってしまう．そのため，MMPC アルゴリズム [3] を用いて因果関係が強く，有向リンクが作成される可能性が高い 2 つのノード間に作られる無向リンクの集合である有向リンク候補を求め，探索対象を有向リンク候補に限定することで，計算量を削減する方法がある．

3　従来手法

BN における構造学習の手法として，Ammar らの手法 [2] がある．Ammar らの手法は，無向グラフに対するアンサンブルを導入することで，真の構造を推定しようとしている．

以下にアンサンブル数を M とした場合の従来アルゴリズムを示す．

Step1) 学習データからブートストラップサンプリングによってサブデータセット $D_i(i = 1, .., M)$ を作成する
Step2) M 個のサブデータセットのそれぞれに対して，無向グラフを推定する
Step3) 各サブデータセットから得られた無向グラフをアンサンブルし，混合無向グラフを求め，有向リンク候補とする
Step4) 有向リンク候補に対して Greedy Hill-Climbing を用いてリンク構造を作成する

3.1　混合無向グラフの作成

Step3 での無向グラフのアンサンブルは，X_p と X_q の間に，m 個のサブデータセットで無向リンクが作成されていた時，閾値 α に対し，$m/M \geq \alpha$ ならば混合無向グラフにおいても X_p と X_q の間に無向リンクを作成する．

3.2　リンク構造の作成

Step4 では，アンサンブルによって作成された混合無向グラフに対して Greedy Hill-Climbing を用いることで，有向リンクを探索的に選択し，リンク構造を作成する．

4　提案手法

Ammar らの手法では，最も重要なリンク構造の推定を，アンサンブル後に全学習データのもとで行う．そのため，アンサンブルは無向グラフの推定にのみ用いられ，リンク構造自体は全学習データから推定される．これは，リンク構造推定時に，学習データに過適合し，ノードの因果関係を適切に推定できないという問題につながる．そこで，提案手法では，各サブデータセットのもとで BN のリンク構造を作成し，これらのリンク構造に対してアンサンブルを行うことで，リンク構造に与えるアンサンブルの効果を大きくし，ノード間の因果関係を適切に表現したリンク構造を推定することを考える．

以下に提案アルゴリズムを示す．

Step1) 学習データからブートストラップサンプリングによってサブデータセット $D_i(i = 1, .., M)$ を作成する

Step2) M 個のサブデータセットのそれぞれに対して，無向グラフを求め，リンク候補を作成する
Step3) 各サブデータセットにおけるリンク候補に対して，Greedy Hill-Climbing を用いてリンク構造を作成する
Step4) 各サブデータセットから作成したリンク構造のアンサンブルによって，混合リンク構造を求める．
Step5) 有向リンクの向きを更新する

4.1 混合リンク構造の作成

Step3 での各サブデータセットにおけるリンク構造の作成において，i 番目のサブデータセットにおけるノード X_p から X_q への向きの有向リンクの有無を $Link_i(p,q)$ で表し，存在すれば 1，存在しなければ 0 と表す．また，**Step4** での混合リンク構造における X_p から X_q への有向リンクの有無を $Link(p,q)$ で表す．$Link(p,q)$ は式 (1) で求められる．

$$Link(p,q) = \begin{cases} 1\ , \omega_{p,q} + \omega_{q,p} \geq \alpha\ \cap\ \omega_{p,q} \geq \omega_{q,p} \\ 0\ , \text{otherwise} \end{cases} \quad (1)$$

$$\omega_{p,q} = \frac{\sum_i Link_i(p,q)}{M} \quad (2)$$

4.2 有向リンクの向きの更新

Step4 の時点では，それぞれの有向リンクの向きの決定を，アンサンブルにおける多数決によって行っている．そこで，**Step5** で混合リンク構造に対して有向リンクの向きを更新する操作を行っている．具体的には有向リンクの反転を行うことで BIC の高くなる任意の有向リンクを 1 つ反転し，どの有向リンクを反転させても BIC スコアが高くならなくならなければ終了する．

5 実験

提案手法の有効性を検証するため，ベンチマークデータである barley（ノード数 48, 有向リンク数 96）を用いて構造学習，確率推論それぞれのステップで評価実験を行った．

5.1 実験 1 構造学習における評価

構造が既知のモデルから学習データを生成し，従来手法，提案手法それぞれを用いて構造学習を行い，式 (3)，式 (4) によって評価した．また，学習データ数 $N = 100, 200$，アンサンブル数 $M = 100, 200$，アンサンブルの閾値 $\alpha = 0.05, 0.10, 0.15, 0.20$ とした．

$$TP = \frac{\text{生成グラフに含まれる正解有向リンクの数}}{\text{正解グラフに含まれる有向リンクの数}} \quad (3)$$

$$FP = \frac{\text{生成グラフに含まれる誤り有向リンクの数}}{\text{生成グラフに含まれる有向リンクの数}} \quad (4)$$

5.2 実験 1 の結果・考察

各条件で 10 回ずつ実験を行った．TP,FP の平均を表 1，表 2 に示す．

表 1. TP による比較

	閾値	アンサンブル数：100		アンサンブル数：200	
		従来	提案	従来	提案
学習データ数：100	0.05	4.18%	9.37%	5.06%	10.00%
	0.10	4.18%	8.10%	5.06%	8.23%
	0.15	4.05%	7.34%	5.06%	7.85%
	0.20	4.05%	7.34%	4.94%	4.94%
学習データ数：200	0.05	15.19%	18.99%	15.06%	18.23%
	0.10	14.81%	16.33%	14.68%	16.20%
	0.15	14.05%	15.44%	13.92%	15.06%
	0.20	13.54%	13.80%	13.67%	14.05%

表 2. FP による比較

	閾値	アンサンブル数：100		アンサンブル数：200	
		従来	提案	従来	提案
学習データ数：100	0.05	38.86%	29.20%	30.05%	23.79%
	0.10	38.86%	23.92%	30.05%	18.55%
	0.15	36.83%	22.66%	27.38%	18.49%
	0.20	35.33%	20.40%	27.88%	15.42%
学習データ数：200	0.05	14.42%	12.32%	15.66%	14.19%
	0.10	14.70%	11.73%	15.90%	14.15%
	0.15	14.68%	9.55%	16.08%	10.31%
	0.20	13.82%	8.37%	14.93%	9.63%

結果より，提案手法で TP が増加し，FP が減少していることから，本研究の目的である，ノードの因果関係を適切に表現したリンク構造を作成できていると考えられ，提案手法の有効性を示すことができた．提案手法では，多様な状況で作成した有向リンクをアンサンブルするため，結びつきが比較的弱いノード間の有向リンクも見つけ出すことができ，TP が改善したと考えられる．

5.3 実験 2 確率推論における評価

48 個のノードのうち，ある 1 つのノードだけを欠損値とし，そのノードの値を，従来手法，提案手法それぞれによって作成したリンク構造を用いて予測する．これをすべてのノードに対して行い，正解率によって評価した．テストデータは各ノードにつき 100 件，閾値 $\alpha = 0.05$ とした．

5.4 実験 2 の結果・考察

各条件で 10 回ずつ実験を行った．正解率の平均を表 3 に示す．

表 3. 正解率による比較

	アンサンブル数：100		アンサンブル数：200	
	従来	提案	従来	提案
学習データ数：100	37.19%	42.02%	38.24%	43.13%
学習データ数：200	44.85%	47.35%	44.82%	47.56%

結果より，提案手法において正解率が改善しており，予測精度についても有効性が確認できた．正解率の改善はノードによって差があったが，正解率が大きく改善しているノードについては，従来手法で表現できていなかった複数のノードと有向リンクで結ばれている構造が確認できた．このようなノードの存在が予測精度向上にも寄与したと考えられる．

6 まとめと今後の課題

本研究では，Ammar らの手法のアルゴリズムにおいて問題となっていた学習データへの過適合に対して，リンク構造のアンサンブルを用いることにより，ノードの因果関係を適切に推定したリンク構造を推定する手法を提案し，その有効性を示した．また，確率推論における予測についても，精度向上が確認された．今後の課題として，学習データに欠損値がある場合への拡張などが挙げられる．

参考文献

[1] 繁枡算男, 本村陽一, 植野真臣,“ベイジアンネットワーク概説,” 培風館, 2006.

[2] Ammar, Sourour, and Philippe Leray. “Mixture of Markov trees for Bayesian network structure learning with small datasets in high dimensional space.” Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Springer Berlin Heidelberg, 2011.

[3]Tsamardinos, Ioannis, Laura E. Brown, and Constantin F. Aliferis. “The max-min hill-climbing Bayesian network structure learning algorithm.” Machine learning 65.1 2006.