データマイニング手法によるベイジアンネットワーク構造学習の高速化

Fast Bayesian Network Learning Algorithm based on Data Mining Technique

森下 民平[*1*2] 植野 真臣[*2]
MORISHITA, Mimpei UENO, Maomi

[*1]株式会社シーエーシー 技術研究グループ
Advanced Technology Group, CAC Corporation

[*2]電気通信大学大学院 情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications

Data mining techniques for frequent itemset mining can improve the performance of calculation of joint probability distributions. We apply the technique to a structure learning algorithm of Bayesian networks which has a desirable property that it leads the structure isomorphic to true causal structure. Our experimental result shows that the proposed algorithm works faster than the original algorithm while keeping accuracy.

1. はじめに

確率的因果構造の表現であるベイジアンネットワークは，推論精度が高いことで知られる．しかしベイジアンネットワークの構造学習は，ノード数に対して NP 完全であること [Chickering 96]，さらにデータ数に対しても，スコアリングベース手法では NP 困難であることが知られており [Chickering 04]，大規模な確率変数を大量データから学習するのは極めて困難な問題である．本研究は，データマイニング手法を組み込むことにより，大規模な確率変数と大量データという困難な状況下でも，効率的に機械学習を行うベイジアンネットワーク構造学習アルゴリズムを提案するものである．

ベイジアンネットワークの推論精度の高さは，複数の先行研究によって明らかにされてきた．たとえば，多数ユーザのアイテム閲覧・購買履歴から，特定ユーザに推薦するアイテムを推薦する協調フィルタリングでは，他手法，すなわち，相関係数法，ベクタ類似度計測法，ベイジアンクラスタリング法のなかで，ベイジアンネットワークに基づく協調フィルタリングの推論精度がもっとも高いことが知られている [Breese 98]．また SIG KDD Cup コンテストのクラス分類問題では，ベイジアンネットワークを用いたクラス分類器が優勝している [Cheng 02b]．

ベイジアンネットワーク構造学習アルゴリズムは，真の同時確率分布とベイジアンネットワークで表現する同時確率分布の近似を最大化することを目的としたスコアリングベース手法と，確率変数間の独立性判定により因果構造を抽出することを目的とした CI (Conditional Independence) ベース，または制約ベースと呼ばれる手法とに大別される．本研究が目的とする大規模な確率変数を扱う場合，既存の AIC [Akaike 74] ，MDL [Rissanen 83] ，UPSM [Cooper 92]，BDeu [Heckerman 95] といったスコアを用いるスコアリングベース手法を採用するより，変数間の独立性判定を行うアプローチを取る方が推論精度は高くなることが，最近 Ueno ら [Ueno 08] によって示された．

しかし，[Ueno 08] が変数間の独立性判定を用いる手法として示したのは，木構造ネットワーク構築手法である MWST [Chow 68] のみであり，真の因果構造が木構造より複雑なネットワークである場合には，真の因果構造を再現することはできない．

本研究では，木構造よりも真の因果構造に近いベイジアンネットワークを学習すれば，推論精度をより向上させられるという仮説のもとに，変数間の独立性判定を行う CI ベース手法の構造学習アルゴリズム TPDA [Cheng 02a] をベースとして用いる．さらに，大規模変数に対する大量データを効率的に処理するために，多頻度アイテム集合を抽出するデータマイニング手法である Apriori [Agrawal 94] をアルゴリズムに組み込み，大規模変数・大量データの状況下でも，より推論精度の高いベイジアンネットワークを効率的に学習するアルゴリズムを提案する．

連絡先: 森下 民平，(株) シーエーシー，東京都中央区日本橋箱崎町 24-1，Phone 03-6667-8070，Fax 03-5641-3213，mimpei@cac.co.jp

2. 関連研究

本研究でベースとする構造学習手法には，大規模変数と大量データへの適応に優れた CI ベースのアルゴリズムのうち，特に大規模変数と大量データに向くことで知られる TPDA を採用する．TPDA 以降も，CI ベースアルゴリズムがいくつか発表されている．PMMS [Brown 05] は TPDA と同じ計算量 $O(N^4)$ ながら TPDA よりも弱い仮定 (DAG faithful assumtion) で動作するよう設計されている．しかしデータ数が大きくなると TPDA より学習時間がかかるため，大規模データには向かない．MBOR [Rodrigues de Morais 08] と PCMB [Peña 07] は 139,351 個の確率変数を扱っているが，ベイジアンネットワークを生成せず，特定のクラス変数に対するマルコフ境界を探すアプローチを取るため，本研究のベースとするにはそぐわない．Kebaili らの手法 [Kebaili 07] は，相関ルールマイニングをベイジアンネットワークの学習と組み合わせている点が本研究との類似点だが，Kebaili らの手法は，CI ベース手法における確率変数間に近似決定的関係や Noisy XOR 関係があるときに正しい因果関係を検出しにくい問題に対処するものであり，実証実験も 6 変数までしかなされておらず本研究とは目的が異なる．

3. ベイジアンネットワーク

ベイジアンネットワークのモデルは，$\mathcal{N} = \langle V, E, \mathcal{P} \rangle$ あるいは $\mathcal{N} = (\mathcal{G}, \mathcal{P})$ と表現される．ここで $\mathcal{G} = (V, E)$ は頂点集合 V，辺集合 E からなる非循環有向グラフで，$\mathcal{P}$ は，$pa(x_n)$ を n 番目の確率変数 x_n の親変数とすると $\mathcal{P} = \{P(x_1|pa(x_1), \ldots, P(x_n|pa(x_n))\}$ で表される条件付確率分布集合である．集合 $\mathcal{P}$ はこれに対応する同時確率分布 $P(\mathcal{U})$ を

$$P(\mathcal{U}) = \prod_{i=1}^{n} P(x_i|pa(x_n)) \tag{1}$$

として与えている．

4. TPDA

TPDA (Three Phase Dependency Analysis) [Cheng 02a] は，一般に CI (Conditional Independence) ベース，あるいは制約ベースと呼ばれる手法に基づくアルゴリズムである．CI ベースアルゴリズムは，データを所与とした確率変数間の条件付き独立テストを行うことにより，データから真の因果構造を推定するアプローチのアルゴリズムのことをいう．

TPDA は，真の因果構造が monotone Dag faithful であると仮定したときに，真の因果構造の推定を保証するアルゴリズムである．Monotone DAG faithful は，DAG faithful よりもやや強い仮定である．ここで，DAG faithful であるとは，真の因果構造が非循環有向グラフであることをいう．$\mathrm{Open}_{\mathcal{N}}(X, Y|\mathbf{C})$ を，モデル $\mathcal{N}$ を所与とした頂点 X，Y のパス上にあって条件として所与としたときに X，Y のパスを開くとき，すなわち互いに依存するときに，所与とする確率変数集合を指すものとする．また $\mathbf{C}$ を所与とした X，Y の条件付き相互情報量を $I(X, Y|\mathbf{C})$ とすると，monotone DAG faithful であるとは，次のように定義される．すなわち，DAG faithful なモデル $\mathcal{N} = \langle V, E, \mathcal{P} \rangle$ は，V の要素であるすべての X，Y について，$\mathrm{Open}_{\mathcal{N}}(X, Y|\mathbf{C}') \subseteq \mathrm{Open}_{\mathcal{N}}(X, Y|\mathbf{C}) \Longrightarrow I(X, Y|\mathbf{C}') \leqslant I(X, Y|\mathbf{C})$ であるときまたそのときに限り，monotone DAG faithful である．

TPDA は，monotone DAG faithful の仮定を用い，条件付き独立テストの際に条件部に与える確率変数集合を順次絞り込むことにより，組み合わせ爆発を回避している．

TPDA で用いる条件付き独立テストでは，以下の条件付き相互情報量を測定し，これが一定の値 ε 以上となる確率変数間にアークをつける．

$$I(X, Y|\mathbf{C}) = \sum_{x,y,\mathbf{c}} P(x, y, \mathbf{c}) \log \frac{P(x, y|\mathbf{c})}{P(x|\mathbf{c})P(y|\mathbf{c})} \tag{2}$$

ここで，大文字 X, Y はそれぞれ確率変数を表し，小文字 x, y は対応する状態値を表す．太字大文字 $\mathbf{C}$ は確率変数集合を表し，太字小文字 $\mathbf{c}$ はそれぞれの変数に対応する状態値集合を表す．

代表的な CI ベースアルゴリズムとして知られる PC アルゴリズム [Spirtes 93] が $O(N^{K+2})$ の計算量 (K は真の因果構造における任意の確率変数の最大次数) であるのに対し，TPDA はノード順を仮定しない場合でも $O(N^4)$ の計算量しか必要とせず，より大規模化するのに適している．CI ベース学習アルゴリズムの実行時間の過半は，変数間の条件付き独立テストの実行に費やされるが，TPDA は実行時間短縮のために，条件付き独立テストの実行回数を少なくするように設計されている．TPDA は，具体的には以下の 3 フェーズを実行し，最後にアークの方向付けを行うことにより構造学習を行う．

(1) Drafting: MWST [Chow 68] を用いて木構造を生成する．すなわち，すべての変数ペアの相互情報量 ((3) 式) を求め，閾値 ε 以上となる変数ペアの間に，無向グラフが閉路を構成しない限りアークを追加する．

$$I(X, Y) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \quad (x \neq y) \tag{3}$$

(2) Thickning: Drafting フェーズで足りない可能性のあるアークをすべて付け加える．すなわち，木構造のうちアークのない変数ペアについて，当該ペアのパス上の隣接ノードを所与とし，(2) 式の条件付き相互情報量を求め，ε 以上であればアークを追加する．

(3) Thinning: ヒューリスティックスを用いて不要なアークを削除し，さらに精密なアーク要否判定を必要とするアークに限り，精密判定を行い不要なアークを削除する．

条件付き独立テストは，条件として与える変数が多くなると指数的に計算負荷が高くなるが，TPDA では，前の段階で条件部に与える変数候補を絞り込むことにより，条件付き独立テストの実行回数を削減している．

TPDA は，既存研究の中ではもっとも大規模変数・大量データに適した構造学習アルゴリズムだが，確率変数の数が多くなると，条件付き独立テストの計算が困難になる．すなわち，条件付き相互情報量 (2) 式右辺のうち，同時確率分布 $P(x, y, \mathbf{c})$ 部分の $\mathbf{c}$ にあたる変数の数が増加するため，計算が困難になる．また同時確率分布のパターンが増えるに従い，計算結果に寄与しない欠損値が多く発生してしまうという問題がある．

5. データマイニングアルゴリズム Apriori

本研究は，前節で示した TPDA の問題を解決し処理を高速化するために，Apriori [Agrawal 94] を提案アルゴリズムの中に組み込む．Apriori は，大規模トランザクションデータに含まれる多頻度アイテム集合を抽出するためのアルゴリズムである．ある集合 A が多頻度アイテム集合でなければ，A を含む集合 B も多頻度アイテム集合ではない，という性質を利用して枝狩りを行うことにより，効率的に多頻度アイテム集合を抽出する．出現頻度の閾値は，最小支持度 δ と呼ぶ同時確率で表される．Apriori の擬似コードを図 1 に示す．

```
Let F_k 長さ k の多頻度アイテム集合
Let D トランザクションデータ集合
Let C_k 長さ k の多頻度アイテム候補集合

 1: F_1 := {x ∈ T|P(x) ≥ δ}
 2: for k := 1 to m do
 3:    C_{k+1} を F_k から生成
 4:    for all t ∈ D do
 5:       if ((c ∈ C_{k+1}) ⊆ t) then
 6:          c.count++
 7:       end if
 8:    end for
 9:    F_{k+1} := {c ∈ C_{k+1}|P(c) ≥ δ}
10: end for
```

図 1: Apriori アルゴリズム

6. 提案アルゴリズム

TPDA の実行時間の 95% は，条件付き独立テストが占め，なかでも条件付き相互情報量 (2) 式における同時確率および条

表 1: 構造学習結果比較 (サンプル数 1,000)

アルゴリズム	M.A.	E.A.	M.O.	W.O.	実行時間
TPDA	9	4	4	2	1.000 (1.000)
提案アルゴリズム ($\delta = 0.001$)	13	7	5	4	0.459 (1.000)
提案アルゴリズム ($\delta = 0.0005$)	10	6	4	2	0.555 (1.000)
提案アルゴリズム ($\delta = 0.0001$)	10	6	4	2	0.558 (1.000)

表 2: 構造学習結果比較 (サンプル数 5,000)

アルゴリズム	M.A.	E.A.	M.O.	W.O.	実行時間
TPDA	8	6	4	1	1.000 (10.279)
提案アルゴリズム ($\delta = 0.001$)	11	7	5	0	0.294 (6.579)
提案アルゴリズム ($\delta = 0.0005$)	9	5	4	1	0.340 (6.290)
提案アルゴリズム ($\delta = 0.0001$)	11	5	4	1	0.297 (5.477)

表 3: 構造学習結果比較 (サンプル数 10,000)

アルゴリズム	M.A.	E.A.	M.O.	W.O.	実行時間
TPDA	11	7	4	3	1.000 (6.254)
提案アルゴリズム ($\delta = 0.001$)	13	7	4	1	0.282 (3.842)
提案アルゴリズム ($\delta = 0.0005$)	11	6	4	0	0.263 (2.956)
提案アルゴリズム ($\delta = 0.0001$)	11	6	4	0	0.347 (3.894)

表 4: 構造学習結果比較 (サンプル数 15,000)

アルゴリズム	M.A.	E.A.	M.O.	W.O.	実行時間
TPDA	9	6	6	4	1.000 (7.641)
提案アルゴリズム ($\delta = 0.001$)	11	6	4	0	0.276 (4.594)
提案アルゴリズム ($\delta = 0.0005$)	12	8	4	1	0.317 (4.354)
提案アルゴリズム ($\delta = 0.0001$)	11	6	4	5	0.388 (5.315)

件付確率を算出するための該当データ数問合せのデータベースクエリーにそのほとんどの時間が費やされている [Cheng 02a]. したがって，条件付き相互情報量計算において，計算に与える影響が小さい部分を枝狩りし，重要部分のみを計算対象として抽出すれば，構造学習の高速化が可能である．本研究の核となるアイデアは，Apriori を用いて最小支持度を満たす同時確率変数とその値の組を高速に抽出し，最小支持度を満たさないものを計算対象から外すことで高速化を達成するというものである．

TPDA 全体を概観すると，後のフェーズに行くにしたがって同時に扱う確率変数の数が増加する傾向にあるアルゴリズムである．したがって，同時に扱う確率変数が順次増加するという特性を持つ Apriori は相性が高く，組み込みやすい．

まず TPDA の第一段階である Drafting フェーズについて考える．Drafting フェーズにおける MWST を用いた木構造の構築では，すべての確率変数ペアについて相互情報量を求め，相互情報量の降順に，閉路を構成しない，すなわち木構造とならない限りはアークを追加していく．このとき TPDA では，相互情報量が ε 未満となる確率変数の組を処理しない．相互情報量は以下の式で表される．

ここで Apriori による同時確率分布計算の枝狩りを行う．すわなち $P(x)$ もしくは $P(y)$ が最小支持度 ($\delta < \varepsilon$) を満たさなければ，$P(x, y)$ も最小支持度を満たさないことに着目し，$P(x)$ もしくは $P(y)$ が最小支持度を満たさなければ，

$$P(x, y) \log \frac{P(x, y)}{P(x)P(y)} < \delta < \varepsilon \tag{4}$$

が成立するため，当該部分の相互情報量計算を省略する．

同様に，次の Thickning フェーズ以降で用いる (2) 式の条件付き相互情報量についても Apriori により計算の枝狩りを行う．(2) 式右辺の $P(x, y, \mathbf{c})$ の因子も，$P(x)$，$P(y)$，$P(\mathbf{c})$，$P(x, y)$，$P(x, \mathbf{c})$，$P(y, \mathbf{c})$ のいずれかが最小支持度 δ より小さければ，$P(x, y, \mathbf{c}) < \delta$ である．したがって，

$$P(x, y, \mathbf{c}) \log \frac{P(a, b|\mathbf{c})}{P(a|\mathbf{c})P(b|\mathbf{c})} < \delta < \varepsilon \tag{5}$$

となるため，当該部分の条件付き相互情報量計算を省略する．

ただし TPDA は，条件付き相互情報量を計算する際，所与とする確率変数集合を大きい方から開始し，順次集合を小さくしていくが，この順序は Apriori による刈り込み順序とは異なる．したがって，より計算を効率化するために，TPDA の一部を改造し，所与とする確率変数集合を小さい方から開始し，順次大きくするように処理順を変更する．

以上のように Apriori により計算が不要な部分を枝狩りし，計算が必要な抽出した部分のみデータベース問合せを実行することにより，構造学習の高速化を行う．

7. 実験結果

TPDA および提案アルゴリズムを実装し，実行時間と正解となる真の因果構造との一致度合いを計測した実験結果を示す．実験として，ベイジアンネットワーク学習の例題として頻繁に用いられる Bayesian Network Repository [BNR] から，Alerm と呼ばれる 37 ノード 46 アークのネットワークを用い，これについてそれぞれ 1,000 件 (表 1)，5,000 件 (表 2)，10,000 件 (表 3)，15,000 件 (表 4) のデータを生成し，各アルゴリズムにより構造学習を行った．各表では，Missing Arc (M.A.), Extra Arc (E.A.), Missing Orientation (M.O.), Wrong Orientation (W.O.), および TPDA による実行時間を 1.000 としたときの実行時間比率を示す．ただし括弧内の実行時間は，当該アルゴリズムを同一パラメータで 1,000 件のデータで実行したときの実行時間を基準とし 1.000 としたときの実行時間比率である．

実験結果より，提案アルゴリズムは TPDA の約 1.8 倍から 3.8 倍程度高速に動作すること，Apriori の最小支持度 δ を小

さくすると構造推定精度を TPDA と同等にできる傾向にあることがわかる．

8. 結論

多頻度アイテム集合抽出を行うデータマイニングアルゴリズム Apriori を同時確率分布計算の枝狩りに用い，確率変数の条件付き独立テストによるベイジアンネットワークを構造学習する TPDA アルゴリズムを高速化する手法を提案した．実験結果により，提案アルゴリズムは，TPDA の構造推定精度をほぼ維持したまま効率的に学習を行えることを示した．

9. 今後の課題

本研究で実装した TPDA は，Cheng ら [Cheng 02a] によるオリジナル実装と比較するとまだ実行速度改善の余地が大きい．データベーススキーマの最適化，クエリーキャッシュ改善などのチューニングを行い，より規模の大きなネットワークを用いた実験を進める予定である．また本研究では，多頻度アイテム集合抽出アルゴリズムとして，古典的で比較的実装が容易な Apriori を用いたが，データ構造として FP-Tree [Han 00] を用いるものなど，より性能の良いアルゴリズムの組み込みも行う予定である．

参考文献

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, in *Proc. of the 20th Int'l Conference on Very Large Databases*, pp. 487–499, Santiago, Chile (1994)

[Akaike 74] Akaike, H.: A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, Vol. 19, pp. 716–723 (1974)

[BNR] Bayesian Network Repository, `http://compbio.cs.huji.ac.il/Repository/`

[Breese 98] Breese, J. S., Heckerman, D., and Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *In proceedgings of Uncertainty in Artificial Intelligence (UAI)*, Vol. 14, pp. 43–52 (1998)

[Brown 05] Brown, L. E., Tsamardinos, I., and Aliferis, C. F.: A Comparison of Novel State-of-the-Art Polynomial Bayesian Network Learning Algorithms, in *Proceedings of The 20th National Conference on Artificial Intelligence (AAAI)*, pp. 739–745, Pittsburgh, Pennsylvania (2005)

[Cheng 02a] Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W.: Learning Bayesian networks from data: an information-theory based approach, *Artificial Intelligence*, Vol. 137, No. 1-2, pp. 43–90 (2002)

[Cheng 02b] Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.-A., Morishita, S., Page, D., and Sese, J.: KDD Cup 2001 Report, *SIGKDD Explorations*, Vol. 3, No. 2, pp. 47–63 (2002)

[Chickering 96] Chickering, D. M.: Learning Bayesian networks is NP-complete, in *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121–130, Springer-Verlag (1996)

[Chickering 04] Chickering, D. M., Heckerman, D., and Meek, C.: Large-Sample Learning of Bayesian Networks is NP-Hard, *J. Mach. Learn. Res.*, Vol. 5, pp. 1287–1330 (2004)

[Chow 68] Chow, C. K. and Liu, C. N.: Approximating discrete probability distributions with dependence trees, *IEEE Trans. on Information Theory*, Vol. 14, No. 3, pp. 462–467 (1968)

[Cooper 92] Cooper, G. F. and Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data, *Machine Learning*, Vol. 9, No. 4, pp. 309–347 (1992)

[Han 00] Han, J., Pei, J., and Yin, Y.: Mining Frequent Patterns without Candidate Generation, in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pp. 1–12, ACM (2000)

[Heckerman 95] Heckerman, D., Geiger, D., and Chickering, D. M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning*, Vol. 20, No. 3, pp. 197–243 (1995)

[Kebaili 07] Kebaili, Z. and Aussem, A.: A novel Bayesian Network structure learning algorithm based on minimal correlated itemset mining techniques, in *Proc. of Second IEEE International Conference on Digital Information Management (ICDIM), Lyon, France*, pp. 121–126, IEEE (2007)

[Peña 07] Peña, J. M., Nilsson, R., Björkegren, J., and Tegnér, J.: Towards scalable and data efficient learning of Markov boundaries, *International Journal of Approximate Reasoning*, Vol. 45, No. 2, pp. 211–232 (2007)

[Rissanen 83] Rissanen, J.: A Universal Prior for Integers and Estimation by Minimum Description Length, *Annals of Statistics*, Vol. 11, pp. 416–431 (1983)

[Rodrigues de Morais 08] Rodrigues de Morais, S. and Aussem, A.: A Novel Scalable and Data Efficient Feature Subset Selection Algorithm, in *ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pp. 298–312, Berlin, Heidelberg (2008), Springer-Verlag

[Spirtes 93] Spirtes, P., Glymour, C., and Scheines, R.: *Causation, Prediction, and Search*, Springer-Verlag, New York, N.Y (1993)

[Ueno 08] Ueno, M. and Yamazaki, T.: Collaborative Filtering for Massive Datasets based on Bayesian Networks, *Behaviormetrika*, Vol. 35, No. 2, pp. 137–158 (2008)