ベイジアン・ネットワーク学習の基礎的性質の分析

植野真臣 電気通信大学　大学院情報システム学研究科

Abstract

近年、ベイジアンネットワークの学習の研究では、ディレクレ事前分布における equivalent sample size (ESS) と呼ばれるハイパーパラメータが非常に重要な役割をはたしていることが報告されている。本研究では、最も一般的なベイジアンネットワークの学習である対数 BDeu (Bayesian Dirichlet equivalence uniform) について漸近分析を行い、ESS の基本的な性質を導く。BDeu は、ESS のアーク追加へのペナルティの働きをする項とアーク追加を助長する項に分解でき、それらがトレードオフの関係になっており、そのために学習効率が ESS に対して非常にセンシティブになることを導いた。ESS の値を大きくするとベイジアンネットワークのアーク数が完全グラフになるまで増え続けるという直感的には逆の結果も証明した。

1　はじめに

近年、ベイジアンネットワークの学習の研究では、ディレクレ事前分布における equivalent sample size (ESS) と呼ばれるハイパーパラメータが非常に重要な役割をはたしていることが報告されている。Steck and Jaakkola (2002) は、ESS が 0 に近づくとアークがつきにくくなること、ESS が大きくなるとアークの数は増加し続けることを示した。Silander, Kontkanen, and Myllymaki(2007) は、シミュレーション実験を行い、Steck らの研究結果を確認した上で、ESS の値に対してベイジアンネットワーク学習が非常にセンシティブであることを示した。しかし、彼らもこれらの現象に対する明確な理由は見つけられなかったと報告している。Ueno (2008) も同様にシミュレーションにより ESS の値に対してベイジアンネットワーク学習が非常にセンシティブであることを示したうえで、経験ベイズによる ESS の最適化法を用いたベイジアンネットワーク学習法を提案している。Steck(2008) は、二ノードにおける BDeu のアークあるなしのベイズファクターの漸近展開が skewness (non-uniformity) と model complexity で表現できることを示している。しかし、この漸近展開では先行研究で示されてきた BDeu における学習の振る舞いを十分には説明できていない。

本研究では、最も一般的なベイジアンネットワークの学習である対数 BDeu (Bayesian Dirichlet equivalence uniform) について漸近分析を行い、ESS の基本的な性質を導く。BDeu は、ESS のアーク追加へのペナルティの働きをする項とアーク追加を助長する項に分解でき、それらがトレードオフの関係になっており、そのために学習効率が ESS に対して非常にセンシティブになることを導いた。ESS の値を大きくするとベイジアンネットワークのアーク数が完全グラフになるまで増え続けるという直感的には逆の結果も証明した。

2　ベイジアンネットワークの学習

N 個の離散変数 $\{x_1, x_2, \cdots, x_N\}$ で、それぞれの値が $\{0, \cdots, r_i - 1\}$ のどれかの状態を取るとする。x_i が k のとき $x_i = k$ と書く。あるベイジアンネットワークの構造 $g \in G$ について, 同時確率分布は以下のように得られる。

$$p(x_1, x_2, \cdots, x_N \mid g) = \prod_{i=1}^{N} p(x_i \mid \Pi_i, g), \tag{1}$$

ここで、 G はすべての可能なベイジアンネットワーク構造集合、Π_i は変数 x_i の親ノード集合を示す。

θ_{ijk} を x_i の親ノードが j 番目のパターンをとったとき ($\Pi_i = j$ と書く) の $x_i = k$ となる条件付き確率パラメータとする。Buntine(1991) は、θ_{ijk} に対し、ディレクレ事前分布を仮定し、EAP (Expected a Priori) 推定値 $\widehat{\theta_{ijk}}$ を用いている。すなわち、

$$\widehat{\theta_{ijk}} = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}}, (k = 0, \cdots, r_i - 2), \tag{2}$$

ここで n_{ijk} は、$\Pi_i = j$ のときの $x_i = k$ となるサンプル数を示し、$n_{ij} = \sum_{k=0}^{r_i-1} n_{ijk}$ とする。α_{ijk} はディレクレ分布のハイパーパラメータを示し、n_{ijk} に対応した疑似サンプルと解釈できる。ここで、$\alpha_{ij} = \sum_{k=0}^{r_i-1} \alpha_{ijk}$, $\widehat{\theta_{ij(r_i-1)}} = 1 - \sum_{k=0}^{r_i-2} \widehat{\theta_{ijk}}$ である。このとき、予測分布は以下のようになる。

$$p(\mathbf{X} \mid g) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{3}$$

ここで、q_i は Π_i のパターン数 $q_i = \prod_{x_l \in \Pi_i} r_l$ を示し、 $\mathbf{X}$ はデータを示す。ベイジアンネットワークの学習とは、上の予測分布を最大化するネットワーク構造を見つけることである。

特に Heckerman *et al.* (1995) は、 likelihood equivalence を満たす十分条件として以下のハイパーパラメータの制約を提案している。彼らはこの基準を BDe(Bayesian Dirichlet equivalence) と呼んでいる。

$$\alpha_{ijk} = \alpha p(x_i = k, \Pi_i = j \mid g^h), \tag{4}$$

ここで α は equivalent sample size (ESS) と呼ばれユーザーが決定できるハイパーパラメータである。g^h はユーザーの事前知識を反映したベイジアンネットワークの仮説構造を示している。

Buntine(1991) が提案した事前分布のハイパーパラメータの制約 $\alpha_{ijk} = \frac{\alpha}{(r_i q_i)}$ は BDe の特別な場合と考えられ、 Heckerman *et al.*(1995) は「BDeu」と呼んでいる。

3 対数予測分布の漸近解析

本節では、対数予測分布の漸近解析により基本的性質を明らかにする。

Theorem 1. 対数予測分布は漸近的に以下により示される。

$$\log p(\boldsymbol{X} \mid g) = \mathcal{H}(g, \alpha) - \mathcal{H}(g, \alpha, \mathbf{X}) - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=0}^{r_i-1} \log \left(1 + \frac{n_{ijk}}{\alpha_{ijk}}\right) \tag{5}$$

ここで $\mathcal{H}(g, \alpha)$,$\mathcal{H}(g, \alpha, \mathbf{X})$ は、以下のエントロピーを示す。

$$\begin{aligned} \mathcal{H}(g, \alpha) &= -\sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=0}^{r_i-1} \alpha_{ijk} \log \frac{\alpha_{ijk}}{\alpha_{ij}} \\ \mathcal{H}(g, \alpha, \mathbf{X}) &= -\sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=0}^{r_i-1} (\alpha_{ijk} + n_{ijk}) \log \frac{(\alpha_{ijk} + n_{ijk})}{(\alpha_{ij} + n_{ij})} \end{aligned}$$

Proof. 式 (3) より対数事後分布は、

$$\begin{aligned} \log p(\mathbf{X} \mid g) &= \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\sum_{k=0}^{r_i-1} \log \Gamma(\alpha_{ijk} + n_{ijk}) - \log \Gamma(\alpha_{ij} + n_{ij}) \right) \\ &\quad + \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\log \Gamma(\alpha_{ij}) - \sum_{k=0}^{r_i-1} \log \Gamma(\alpha_{ijk}) \right) \end{aligned}$$

a が十分大きい時の以下のスターリン展開 [2] を用いる。

$$\log \Gamma(a) = \frac{1}{2} \log(2\pi) + \left(a - \frac{1}{2}\right) \log a - a + \mathcal{O}\left(\frac{1}{a}\right),$$

$\alpha + n$ が十分大きいとき、

$$\begin{aligned}
& \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\sum_{k=0}^{r_i-1}\log\Gamma(\alpha_{ijk}+n_{ijk})-\log\Gamma(\alpha_{ij}+n_{ij})\right) \\
= & \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\sum_{k=0}^{r_i-1}(\alpha_{ijk}+n_{ijk})\log(\alpha_{ijk}+n_{ijk})-(\alpha_{ij}+n_{ij})\log(\alpha_{ij}+n_{ij})\right. \\
& \left.+\frac{r_i-1}{2}\log(2\pi)-\frac{1}{2}\sum_{k=0}^{r_i-1}\log(\alpha_{ijk}+n_{ijk})+\frac{1}{2}\log(\alpha_{ij}+n_{ij})\right)+\mathcal{O}(\frac{\sum_{i=1}^{N}r_iq_i}{n+\alpha}) \\
= & \sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}(\alpha_{ijk}+n_{ijk})\log\frac{(\alpha_{ijk}+n_{ijk})}{(\alpha_{ij}+n_{ij})} \\
& +\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\frac{r_i-1}{2}\log(2\pi)-\sum_{k=0}^{r_i-1}\log(\alpha_{ijk}+n_{ijk})+\log(\alpha_{ij}+n_{ij})\right)+\mathcal{O}(\frac{\sum_{i=1}^{N}r_iq_i}{n+\alpha})
\end{aligned}$$

同様に

$$\begin{aligned}
& \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\log\Gamma(\alpha_{ij})-\sum_{k=0}^{r_i-1}\log\Gamma(\alpha_{ijk})\right)=-\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\alpha_{ijk}\log\frac{\alpha_{ijk}}{\alpha_{ij}} \\
& -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\frac{r_i-1}{2}\log(2\pi)-\sum_{k=0}^{r_i-1}\log\alpha_{ijk}+\log\alpha_{ij}\right)+\mathcal{O}(\frac{\sum_{i=1}^{N}r_iq_i}{n+\alpha})
\end{aligned}$$

従って

$$\begin{aligned}
\log p(\mathbf{X}\mid g) &= \mathcal{H}(g,\alpha)-\mathcal{H}(g,\alpha,\mathbf{X})-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\sum_{k=0}^{r_i-1}\log(\alpha_{ijk}+n_{ijk})-\log(\alpha_{ij}+n_{ij})\right) \\
&\quad +\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left(\sum_{k=0}^{r_i-1}\log\alpha_{ijk}-\log\alpha_{ij}\right)+\mathcal{O}(\frac{\sum_{i=1}^{N}r_iq_i}{n+\alpha}) \\
&= \mathcal{H}(g,\alpha)-\mathcal{H}(g,\alpha,\mathbf{X})-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\log\left(1+\frac{n_{ijk}}{\alpha_{ijk}}\right) \\
&+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\log\left(1+\frac{n_{ij}}{\alpha_{ij}}\right)+\mathcal{O}(\frac{\sum_{i=1}^{N}r_iq_i}{n+\alpha}).
\end{aligned}$$

N が十分に大きい時、$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\log\left(1+\frac{n_{ijk}}{\alpha_{ijk}}\right)+\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\log\left(1+\frac{n_{ij}}{\alpha_{ij}}\right)$ は、$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\log\left(1+\frac{n_{ijk}}{\alpha_{ijk}}\right)$ に収束する。

□

この定理より、以下の性質が簡単に証明される。

Corollary 1. *For* $\forall i,\forall j,\forall k$, $\alpha_{ijk}=\frac{1}{3}n_{ijk}$、対数予測分布は *AIC [1]* と事前分布エントロピーの和として表現される。

$$\begin{aligned}
\log p(\boldsymbol{X}\mid g) &= \mathcal{H}(g,\alpha)+l(\widehat{\theta}\mid\mathbf{X})-\sum_{i=1}^{N}q_i(r_i-1) \\
&\quad \mathcal{H}(g,\alpha)-AIC
\end{aligned}$$

このことより、AIC 最小化によるベイジアンネットワーク学習は、事前知識を一切用いず、事前分布を経験的に与える対数予測分布に一致することがわかる。

Corollary 2. *For* $\forall i, \forall j, \forall k$, $\alpha_{ijk} = 1$*(*事前分布が一様分布*)*、n が十分に大きい時、対数予測分布の上限は *BIC [6]* と事前分布エントロピーの和に収束する。

$$\begin{aligned} \log p(\boldsymbol{X} \mid g) &\approx \mathcal{H}(g, \alpha) - \mathcal{H}(g, \alpha, \mathbf{X}) - \frac{1}{2} \sum_{i=1}^{N} q_i r_i \log(1 + n_{ijk}) \\ &\to \mathcal{H}(g, \alpha) - BIC \end{aligned}$$

一様分布を事前分布に与えた場合の対数予測分布は BIC とほぼ同等の振る舞いをすることがわかる。

4 BDeu の漸近解析

事前の知識を事前分布に与えることは難しいので、一般には、事前分布のハイパーパラメータの制約 $\alpha_{ijk} = \frac{\alpha}{(r_i q_i)}$ を持つ BDeu が用いられることが多い。ここでは、定理 1 より、log-BDeu の性質を解析することにする。以下の性質が成り立つ。

Corollary 3. $\alpha + n$ が十分に大きい時、*Log-BDeu* は、以下に収束する。

$$\log p(\boldsymbol{X} \mid g) = \alpha \sum_{i=1}^{N} \log r_i - \mathcal{H}(g, \alpha, \mathbf{X}) - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=0}^{r_i - 1} \log \left(1 + \frac{r_i q_i n_{ijk}}{\alpha}\right) \tag{6}$$

情報量基準として解釈した場合、$\alpha \log r_i$ は親ノード数に対して定数であり無視できる。結果として、log-BDeu は次の二つに分解できる。(1) 対数事後分布 $-\mathcal{H}(g, \mathbf{X})$ と (2) パラメータ数へのペナルティ項 $\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=0}^{r_i - 1} \log \left(1 + \frac{(r_i q_i n_{ijk})}{\alpha}\right)$。これは、モデルへのあてはまりを反映した (1) 項とアークの追加を阻止するためのペナルティ項 (2) よりなるよく知られたモデル選択基準の形式に一致する。

Steck (2008) も二ノードのアークのありなしに関する対数ベイズファクターの漸近展開により、それがデータ分布の非一様性とモデルの複雑さとのトレードオフになることを導出している。しかし、この分析では、先行研究で示されてきた α についての BDeu の振る舞いを説明できていない。具体的には、Steck (2008) の導出では、ベイジアンネットワーク学習における α とデータ数との関係がほとんど考慮されていない。

本結果では、対数事後分布における α は、経験エントロピー（一様性）を増加させ、アークの追加を阻止する働きをする。一方、ペナルティ項での α は単調にペナルティ項の値を減少させ、アークの追加を増加させる働きをする。すなわち、α は対数事後分布とペナルティ項で全く逆の働きをしているのである。言葉を変えると、本結果は対数事後分布における α とペナルティ項における α のアーク追加と削除の働きによるトレードオフの関係があることを示している。最も重要なことは、この α のトレードオフが、先行研究で問題となってきた BDeu の α への高いセンシティビティの原因となっていると考えられることである。

さらに以下の性質が証明できる。

Proposition 1. $\alpha \to 0$ のとき*, BDeu* により学習されるネットワーク構造のアーク数は *0* に近づく。

証明は Corollary3 より明らかである。また、この結果は、Steck & Jaakkolla (2002) で示されたシミュレーションの結果と同じである。

Proposition 2. $\alpha \to \infty$ のとき*, BDeu* により学習されるネットワーク構造のアーク数は単調増加し、完全グラフを得る。

Proof. $\alpha \to \infty$ のとき、ペナルティ項 $-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\sum_{k=0}^{r_i - 1} \log(1 + \frac{(r_i q_i n_{ijk})}{\alpha}) \right) + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\log(1 + \frac{(r_i q_i n_{ij})}{\alpha}) \right)$ は 0 に収束する。従って、log-BDeu は以下に収束する。

$$\log p(\mathbf{X} \mid g) = \alpha \sum_{i=1}^{N} \log r_i - \mathcal{H}(g, \mathbf{X})$$

対数事後分布はアークの増加に対して単調増加するので log-BDeu も単調増加する。

□

Table 1: α の変化による BDeu の振る舞い

	N	r_i		α	0.1	1	10	100	1,000	10^4	10^5	10^6
alarm	36	4	BDeu	+	3.0	3.6	32.0	123.4	192.7	254.0	314.4	370.0
				-	1.6	1.3	0.4	0.1	0.4	0.4	0.4	0.0
				ME	4.7	4.9	32.3	123.5	192.1	254.0	314.8	370.0
				SD	1.66	1.57	2.81	3.41	3.40	2.66	2.84	2.67
insurance	27	5	BDeu	+	1.0	0.9	9.8	45.8	91.9	132.6	173.0	205.4
				-	15.0	12.5	9.9	8.0	7.4	7.4	7.4	7.4
				ME	16.0	13.4	19.6	53.8	99.3	139.8	180.4	212.8
				SD	1.70	1.11	1.06	2.06	2.54	2.16	2.64	3.16
water	32	4	BDeu	+	15.9	21.3	51.2	104.3	152.2	193.9	236.5	275.4
				-	40.2	36.6	27.9	20.3	17.1	14.3	12.6	10.9
				ME	56.1	57.8	79.1	124.6	169.3	208.1	249.1	286.3
				SD	1.60	1.98	3.18	2.95	3.42	3.28	2.58	3.12
win95pts	76	2	BDeu	+	51.3	80.0	218.1	447.5	665.3	879.4	None	None
				-	17.5	16.9	12.3	9.7	6.1	5.8	None	None
				ME	68.8	96.9	230.5	457.2	671.4	885.2	None	None
				SD	10.48	10.62	8.11	6.55	5.62	5.23	None	None

5 数値例

ここでは、シミュレーション実験によって α の値を変化させて BDeu の振る舞いを分析する。本実験では、Bayesian Network Repository (http://compbio.cs.huji.ac.il/Repository/) より、Alarm network, Insurance network, Water network, Win95pts network の構造を用いて実験を行った。実験は各構造より 1,000 の乱数データを発生させ、α の値を変化させながら BDeu で学習を各 50 回繰り返し行い、平均誤差を評価した。表 1 に結果を示した。「ME」は 50 回中で学習された余分なアークと引かれなかった真のアークの和の平均値を示しており、「SD」は ME の標準偏差を示している。「+」は余分なアークの平均数を示し、「-」は引かれなかった真のアークの平均数を示す。None は、計算時間が 3 日間を超え、解が得られなかったことを示している。表より、α の値が増えるとアーク数が単調増加していること、α の値を減少させていくとアークに対するペナルティが強くなり、アーク数が減少していくことがわかる。すなわち、Proposition 1 や Proposition 2 が確認された。

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** (6): 716–723.

[2] Box,G.E.P. & Tiao,G.C. (1992). Bayesian Inference in Statistical Analysis. New York, N.Y.: John Wiley and Sons, Inc.

[3] Buntine, W. L. (1991) Theory refinement on Bayesian networks. In B. D'Ambrosio, P. Smets and P. Bonissone (eds.), *Proc. of the 7-th Int. Conf. Uncertainty in Artificial Intelligence*, pp. 52–60. Morgan Kaufmann Publishers.

[4] Lam, W. & Bacchus, F. (1994) Learning Bayesian Belief Networks: An Approach Based on the MDL Principle. *Computational Intelligence* **10**(4):269-293.

[5] Rissanen, J. (1978) Modeling by shortest data description. *Automatica* **14**:465–471.

[6] Schwarz, G.E. (1978) Estimating the dimension of a model, *Annals of Statistics* **6**(2):461–464.

[7] Silander, T., Kontkanen, P. & Myllymaki, P. (2007) On sensitivity of the MAP Bayesian network structure to the equipment sample size parameter, In K.B. Laskey, S.M. Mahoney and J. Goldsmith (eds.), *Proc. the 23d conference of Uncertainty in Artificial Intelligence*, pp. 360–367, Morgan Kaufmann Publishers.

[8] Steck, H. & Jaakkola, T.S. (2002) On the Dirichlet Prior and Bayesian Regularization. *Advances in Neural Information Processing Systems (NIPS)*. MIT Press.

[9] Steck, H. (2008) Learning the Bayesian network structure: Dirichlet Prior versus Data. In D.A. McAllester and P. Myllymaki (eds.), *Proc. the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 511–518. Morgan Kaufmann Publishers.

[10] Suzuki, J. (1993) A Construction of Bayesian networks from Databases on an MDL Principle, In D. Heckerman and E. H. Mamdani (eds.), *Proc. the 9th conf. Uncertainty in Artificial Intelligence*, pp. 266–273, Morgan Kaufmann Publishers.

[11] Suzuki, J. (1999) Learning Bayesian Belief Networks based on the Minimum Description Length Principle: Basic Properties. *IEICE Trans. on Fundamentals* **E82**-A:2237–2245.

[12] Bouckaert, R. (1994) Probabilistic network construction using the minimum description length principle. *Technical Report ruu-cs-94-27*, Utrecht University.

[13] Heckerman, D., Geiger, D., & Chickering, D. (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* **20**:197–243.

[14] Maomi Ueno (2008) Learning likelihood-equivalence Bayesian networks using an empirical Bayesian approach, *Behaviormetrika*, Vol. 35, No. 2, 115–135

[15] Yang, S. & Chang, K-C. (2002) Comparison of score metrics for Bayesian network learning. *IEEE Transaction on systems, Man and Cybernetics – Part A: Systems and Humans* **32**(3):419–428.