2D1-1

ベイジアン・ネットワークにおける情報量規準

Information Criterion applied to Bayesian Network

植野 真臣*1 久保 治彦*1 山崎 敬広*2
Maomi Ueno Haruhiko Kubo Takahiro Yamazaki

*1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.

*2日本電信電話 (株) 未来ねっと研究所
NTT Network Innovation Laboratories

Learning Bayesian networks based on MDL (Minimum Description Length) is an approximation method of an exact Bayesian predictiove distribution, which is called "DPSM(Dirichlet Prior Scoring Metrics)". However, some experimental researches showed that MDL based learning is less effective than DPSM based learning. The main reason of this is that DPSM can reflect the prior distribution but MDL can not reflect it. Therefore, this paper proposes a MDL which reflects user's prior knowledge. The unieque features of the MDL are as follows: 1. It is an approximation of the predictive distribution based on the Dirichret-multinomial model for Bayesian networks, and 2. It has a strong consistency for any prior knowkledge and any hyper-parameter. Some simulation experiments show the effectiveness of the proposed model.

1. はじめに

現在用いられるベイジアン・ネットワーク学習の Scoring Metrics は大別して、1) ベイジアン予測分布としての DPSM(Dirichlet Prior Scoring Metrics) [1] と 2)DPSM におけるハイパーパラメータをある値に固定したときに近似して導かれる MDL(Minimum Description Length) [3]、[4] に分類される。しかし、多くの研究（たとえば、Yang 02 [5]）で、DPSM の予測精度に対して MDL が非常に劣っているという結果が報告されている。この根拠は、DPSM が事前分布を反映しているのに対し、MDL が事前分布を全く反映できないことにあると考えられる。Ueno (2008) [6] は、これまで DPSM において様々な最適なハイパーパラメータの存在が過去の研究ごとに主張されてきたのに対し、どのようなハイパーパラメータ（構造に関する事前知識）を設定しても漸近一致性が成り立つことを数学的に証明し、事前分布におけるハイパーパラメータを経験ベイズにより推定することにより、これまで存在するベイジアン・ネットワーク学習の予測精度を大幅に向上できることを示している。このとき、経験ベイズにより求められる最適なハイパーパラメータの値は、データの種類、データ数などで様々に変化することも示している。すなわち、予測に最適な事前知識はデータに応じてすべて異なることになり、如何にデータごとに妥当な事前知識を設定できるかがベイジアン・ネットワーク学習の改善に重要であると考えられる。

本論では、MDL に基づく学習効率の向上のために事前知識を反映する MDL_{prior} の提案を行う。具体的には、ハイパーパラメータの値を固定しない状態で DPSM の対数をスターリング展開し MDL を導く。導出された MDL の性質は以下のとおりである。

1. ベイズ理論に忠実に事前知識を反映することができる
2. どのような事前知識に対しても漸近一致性を持つ
3. 漸近的には従来 MDL と DPSM と同じふるまいをする
4. どのような事前知識に対しても Likelihood equivalence を持つ

シミュレーション実験を行い、以下の知見を得た。

1. 真の構造に近い事前知識を与えた場合、少数データでも学習効率を向上させることができる。
2. 全く真の構造とは異なる事前知識を与えた場合でも、データ数が増えると学習効率が向上する。また、この場合でも従来よりも学習効率が高い。
3. 構造に関する事前知識がない場合、一様分布を設定することができ、この場合でも従来の MDL よりも学習効率が良い。

: 電気通信大学大学院 情報システム学研究科, 〒 182-8585
東京都調布市調布ヶ丘 1-5-1

2. ベイジアン・ネットワーク

今, $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ を離散 N 変数集合とし, 各変数は r_i 個の状態集合 $\{0, \cdots, r_i - 1\}$ の中からひとつの値をとるとする. ここで, 変数 x_i が値 k をとるときに $x_i = k$ と書くこととし, ベイジアンネットワークは確率構造 S と条件付き確率パラメータ集合 Θ によって (S, Θ) として表される. 図 1 は確率構造 S の一例である.

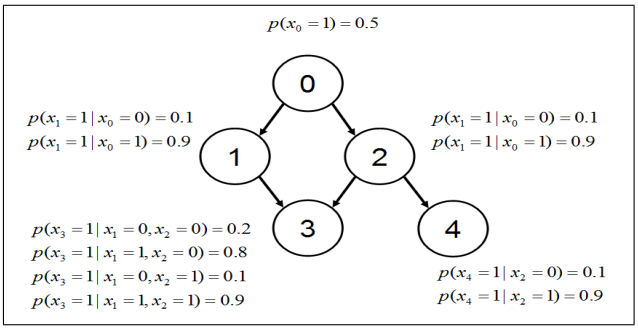

図 1: ベイジアンネットワークの図例

ベイジアンネットワークでは, 構造 S を所与, 変数 i の親ノード集合を $\Pi_i \subseteq \{x_1, \cdots, x_N|\}$ として, 同時確率分布を以下の式 (1) のように表わすことができる.

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

今, θ_{ijk} を親ノード変数集合 Π_i が j 番目のパターンをとったときの $x_i = k$ となる条件付き確率を示すパラメータとする. このとき, データ $\boldsymbol{X}$ を得たときの事後確率は, 事前分布にディレクレィ分布を考えて以下のとおりとなる。

$$p(\Theta_{B_S} \mid \mathbf{X}, B_S) \propto \prod_{i=1}^{N} \prod_{j=1}^{q_i} \prod_{k=0}^{r_i-1} \theta_{ijk}^{n'_{ijk}+n_{ijk}-1}. \tag{2}$$

この事後分布における推定値は

$$\widehat{\theta_{ijk}} = \frac{n'_{ijk} + n_{ijk}}{n'_{ij} + n_{ij}}, (k = 0, \cdots, r_i - 2), \tag{3}$$

ここで $n'_{ij} = \sum_{k=0}^{r_i-1} n'_{ijk}, n_{ij} = \sum_{k=0}^{r_i-1} n_{ijk}, \widehat{\theta_{ij(r_i-1)}} = 1 - \sum_{k=0}^{r_i-2} \widehat{\theta_{ijk}}$.

これより予測分布を求めると

$$p(\mathbf{X} \mid B_S) \propto \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(n'_{ij})}{\Gamma(n'_{ij} + n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(n'_{ijk} + n_{ijk})}{\Gamma(n'_{ijk})} \tag{4}$$

が得られ、DPSM と呼ばれることが多い [1].

3. MDL

一方, Learning Bayesian netowrk における MDL (Minimum Description Length) の適用は、Lam [2] の研究に始まるが、それが一致性を持っていないことより、Bouckaert [3] や Suzuki [4] により厳密な予測分布 DPSM(4) の近似として以下のよく知られた MDL 式が導かれている。

$$\sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=0}^{r_i-1} -n_{ijk} \ln \frac{n_{ijk}}{n_{ij}} + K \ln n_{ij} \tag{5}$$

ここで、K はモデルのパラメータ数を示す。ただし、この数式は Bouckaert はベイジアンネットワークのディレクレ - 多項モデルにおいて事前分布に一様分布を仮定して導いたものであり、Suzuki は Bouckaert の展開は間違いでハイパーパラメータの値が 1/2 となるときに上式が導かれると主張している。最近では、Ueno [6] は、どのようなハイパーパラメータの値についても漸近的に式 (4) が式 (5) に収束することを証明している。このことは,Bouckaert[3]、Suzuki[4] の導出が共に正しかったことを示した。

Yang ら [5] はシミュレーション実験で、ディレクレ - 多項予測分布や K2,BDe などの他の Scoring metrics と比較して上の MDL が精度が悪いことを報告している。このことは、他の Scoring metrics に比較して、先の MDL が事前知識を反映できないために、たとえば、データ数が少ない時には $ln\frac{n_{ijk}}{n_{ij}} + K \ln n_{ij}$ が 0 になったりしてしまうことに影響していると考えられる。

そこで本論では、Ueno [6] の展開の一部を用いて、事前知識を反映した MDL の導出を行う。

4. 事前知識を反映した MDL

本論では, 予測分布 (4) の対数において、先行研究 [3]、[4] が行っているようなハイパーパラメータを固定することはせず、直接以下式をスターリング展開により以下の MDL を導くことができる。

$$\begin{aligned}
&\ln p(\mathbf{X} \mid B_S) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\sum_{k=0}^{r_i-1} \ln \Gamma(n'_{ijk} + n_{ijk}) - \ln \Gamma(n'_{ij} + n_{ij}) \right) + const.
\end{aligned}$$

スターリング展開を用いて 以下が得られる。

$$\begin{aligned}
&\ln p(\mathbf{X} \mid B_S) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\begin{array}{l} \frac{r_i-1}{2} \ln(2\pi) + \sum_{k=0}^{r_i-1} \left(n'_{ijk} + n_{ijk} - \frac{1}{2} \right) \\ \ln(n'_{ijk} + n_{ijk}) \\ -\left(n'_{ij} + n_{ij} - \frac{1}{2} \right) \ln(n'_{ij} + n_{ij}) \end{array} \right) \\
&\quad + const, (n \to \infty) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\begin{array}{l} \sum_{k=0}^{r_i-1} (n'_{ijk} + n_{ijk}) \ln(n'_{ijk} + n_{ijk}) \\ -(n'_{ij} + n_{ij}) \ln(n'_{ij} + n_{ij}) + \frac{r_i-1}{2} \ln(2\pi) \\ -\frac{1}{2} \sum_{k=0}^{r_i-1} \ln(n'_{ijk} + n_{ijk}) + \frac{1}{2} \ln(n'_{ij} + n_{ij}) \end{array} \right) \\
&\quad + const, (n \to \infty).
\end{aligned}$$

$\ln(n'_{ij} + n_{ij}) \geq \ln(n'_{ijk} + n_{ijk})$ なので, 以下が導ける。

$$\begin{aligned}
&\ln p(\mathbf{X} \mid B_S) \\
&\geq \sum_{i=1}^{N} \sum_{j=1}^{q_i} \left(\begin{array}{l} \sum_{k=0}^{r_i-1} (n'_{ijk} + n_{ijk}) \ln \frac{(n'_{ijk} + n_{ijk})}{(n'_{ij} + n_{ij})} \\ -\frac{r_i - 1}{2} \ln \frac{(n'_{ij} + n_{ij})}{2\pi} \end{array} \right) \\
&\quad + const, (n \to \infty).
\end{aligned}$$

$\ln(n' + n) \geq \ln(n' + n_{ij})$ より、

$$\begin{aligned}
\ln p(\mathbf{X} \mid B_S) \quad &\geq \quad \ln p(\mathbf{X}, \hat{\Theta}_{B_S} \mid B_S) - \left(\frac{K}{2} \right) \ln \frac{n' + n}{2\pi} \\
&\quad + const
\end{aligned} \tag{6}$$

を得る。ここで、K はモデルのパラメータ数を示す。

これは、式 (5) 同様に強一致性を持つ MDL となるが, 従来の MDL と異なることは、数式中にハイパーパラメータ n'_{ijk} が残っていることに注意してほしい。

このことは、ハイパーパラメータによって表現された事前知識の影響を反映することを示している。そこで、式 (6) の右辺を MDL とし、うまく事前知識を与えることにより、学習効率をより高くすることできると考えられる。

従って、事前知識を反映する MDL を MDL_{prior} として以下のように定義する。

$$MDL_{prior} = \ln p(\mathbf{X}, \hat{\Theta}_{B_S} \mid B_S) - \left(\frac{K}{2} \right) \ln \frac{n' + n}{2\pi} \tag{7}$$

さらに

$$MDL_{prior} \to_{n \to \infty} MDL + const \tag{8}$$

となり、漸近的には、MDL_{prior} はどのような事前知識に対しても従来の MDL や DPSM と同様の振る舞いをする。

また、MDL_{prior} に事前知識を与えやすいように BDe メトリック同様に Likelihood Equivalence を満たすように以下のような事前知識をハイパーパラメータに反映させる.

$$n'_{ijk} = p(x_i = k, \prod_i = j|S)n' \tag{9}$$

ここで、$p(x_i = k, \prod_i = j|S)$ は事前知識となる確率ネットワークの同時確率分布を示している。一般には事前知識を用いてネットワーク構造を構築する。多くの場合、条件付き確率パラメータは、一律に $1/(r_i - 1)$ や 0.8,0.2 の組を与えたりする。

5. シミュレーション実験と結果

5.1 実験

図 1 の構造よりデータを 100 個,500 個,1000 個,10000 個発生させ, Greedy Serch を用いた MDL_{prior} および従来の MDL により 11 種類の事前知識となる確率ネットワーク構造をそれぞれに与え構造学習を行う. 事前知識となる構造は、真の構造のアークのないところにいくつかのアークを加える、アークのあるところにいくつかのアークを削除する、という操作を行って、真の構造と生成される構造の異なるアーク数を 0,1,2 と増やしながら構造を作成した。事前分布の同時確率の計算にはモンテカルロ法を用いた近似を行った.

5.2 結果

上の実験で、ハイパーパラメータは 1 から 100 まで変化させ, このプロセスをそれぞれ 1000 回繰り返した結果, MDL_{prior} について正しく構造を推定した回数をデータ数ごとに図 2 − 図 5 に示す. 図中の縦軸は 1000 回中モデルが的中した回数を, 横軸はハイパーパラメータの値を示す. また、各結果は、上から、真の構造と生成される構造の異なるアーク数が 0,1,2 と変化していったときの結果を示している。

これらの結果より, 真の構造に近い事前知識を与えた場合、特に少数データからの学習で推定精度に効果があることがわかる. 一方、誤った構造を事前知識に与えたときにも、漸近一致性によりデータ数が増えれば推定精度は真の構造に近い事前知識を与えた場合に近くなり、10,000 のデータでは事前知識（ハイパーパラメータの値）に関係なく、ほぼ 100 %の確率で真の構造を推定することがわかる.

表 1 に MDL_{prior} と従来の MDL でのサンプル数を変化させた時の学習効率を示し、+は推定した構造のアークで真の構造に含まれないときの数、-は推定した構造でアークがないのに真の構造ではアークが存在しているときの数、○は 1000 回中、正確に真の構造を推定できた回数を示す。

MDL_{prior}（一様分布）は、特に事前知識がない場合、Likelihood equivalence を満たすように $n'_{ijk} = n'/(r_i q_i)$ と一様分布となるようにハイパーパラメータを設定したときの結果を、MDL_{prior}（最高成績構造）は、事前知識となる構造の中で最も成績の良かった構造 (真の構造) を事前知識として用いたときの結果、MDL_{prior}（最低成績構造) は、事前知識となる構造の中で最も成績の悪かった構造（真の構造から最もかい離した構造）を事前知識として用いたときの結果を示している。

結果より、少数データのとき、正しい事前知識を持っている場合、圧倒的に学習効率が良いことがわかる。興味深いのは、事前知識がない場合でも、従来の MDL に比較してよい学習効率を示している点、また、間違った事前知識を用いても、データ数が５００－１０００で従来のＭＤＬよりも学習効率が良くなっている点にある。MDL_{prior} では、条件付き確率パラメータを推定するときに、ハイパーパラメータ n'_{ijk} を用いることができるために（予測に効果がある）縮約統計量としての効果があったと解釈できる。また、今回、小さいネットワークで行っているが、大きなネットワークでは、同時パターンの数が指数的に増え、missing data の数が増えてしまうので、従来の MDL の学習成績はより悪くなると考えられる。

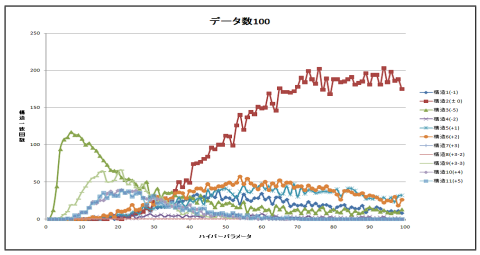

図 2: 様々な事前知識を与えたときの学習効率 (データ数 100)

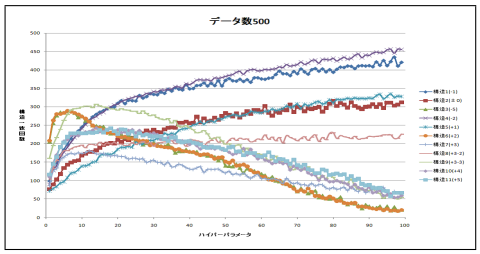

図 3: 様々な事前知識を与えたときの学習効率 (データ数 500)

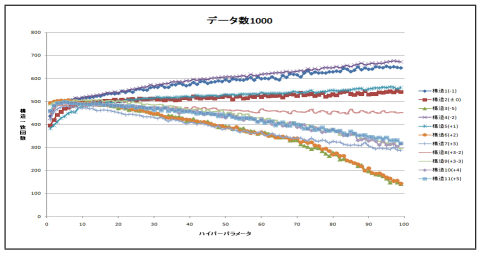

図 4: 様々な事前知識を与えたときの学習効率 (データ数 1000)

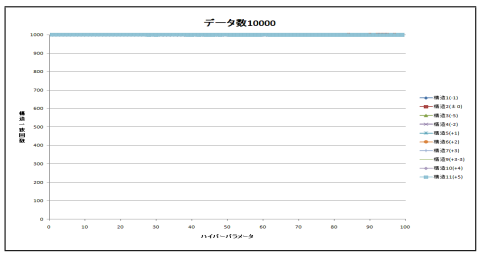

図 5: 様々な事前知識を与えたときの学習効率 (データ数 10000)

6. 事前知識を組み込んだベイジアン・ネットワーク・ソフトウェア

著者がこれまで開発してきたベイジアン・ネットワーク・ソフトウェア ”Bayesian Discovery”に以上の MDL_{prior} を搭載し、ユーザーの事前知識を構造として入力することができるように開発を行った。

7. 終わりに

本論では、MDL に基づく学習効率の向上のために事前知識を反映する MDL_{prior} の提案を行った。導出された MDL の性質は以下のとおりである。

表 1: 事前知識を反映する MDL と従来 MDL との学習効率の比較

	従来の MDL			MDL_{prior}（一様分布）			MDL_{prior}（最高成績構造）			MDL_{prior}（最低成績構造）		
n	+	-	○	+	-	○	+	-	○	+	-	○
100	47	1046	6	2311	534	14	94	654	323	3676	332	1
500	7	915	85	664	592	214	31	540	456	490	710	180
1000	10	779	221	245	307	586	18	315	676	131	472	479
10000	0	0	1000	0	0	1000	0	0	1000	0	0	1000

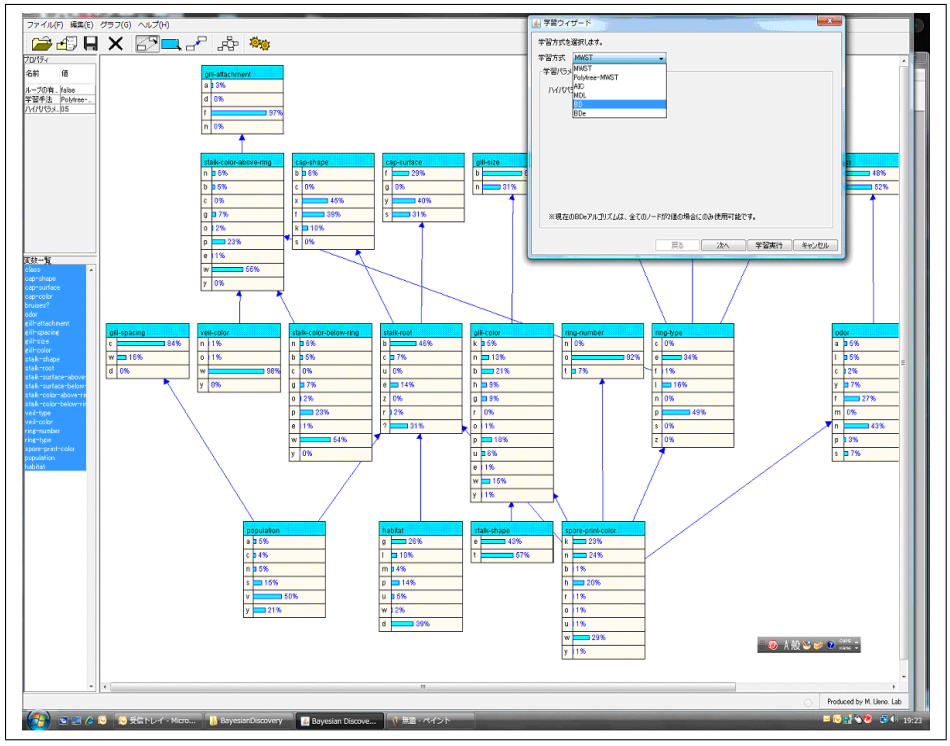

図 6: 事前知識を組み込んだベイジアン・ネットワーク・ソフトウェア ”Bayesian Discovery”

1. ベイズ理論に忠実に事前知識を反映することができる
2. どのような事前知識に対しても漸近一致性を持つ
3. 漸近的には従来 MDL と DPSM と同じふるまいをする
4. どのような事前知識に対しても Likelihood equivalence を持つ

シミュレーション実験を行い、以下の知見を得た。

1. 真の構造に近い事前知識を与えた場合、少数データでも学習効率を向上させることができる。
2. 全く真の構造とは異なる事前知識を与えた場合でも、データ数が増えると学習効率が向上する。また、この場合でも従来よりも学習効率が高い。
3. 構造に関する事前知識がない場合、一様分布を設定することができ、この場合でも従来の MDL よりも学習効率が良い。

参考文献

[1] Cooper,G.F. and Herskovits,E.: A Bayesian method for the induction of probabilistic networks from data, Machine Learning, 9, pp.54-62, (1992)

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

[3] Bouckaert, R. :Properties of Bayesian network learning algorithm, *Proc. Uncertainty in Artificial Intelligence, California*, pp.102-109. (1994).

[4] Suzuki, J.: Learning Bayesian belief networks based on the MDL principle:An efficient algorithm using the branch and bound technique, *IEICE Transction, Information and Systems*, Vol.**E81-D**, No.12. ,pp.356-367 (1998).

[5] Yang, S. and Chang, K-C.: Comparison of score metrics for Bayesian network learning, *IEEE Transaction on systems, Man and Cybernetics-PART A: Systems and Humans*, Vol.**32**, NO. 3, 419-428. (2002).

[6] Ueno, M. : Learning likelihood-equivalence Bayesian networks using an empirical Bayesian approach, *Behaviormetrika*, Vol. **35**, No.1, (2008)