経験的 BDeu の最適化によるベイジアン・ネットワークの学習

Learning Bayesian Network by optimization of empirical BDeu

大川淳史*1　植野真臣*1
Atsushi Okawa　Maomi Ueno

*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

This paper researches the validity of a hypothesis that an ESS(Equivalent Sample Size) value to maximize the predictive distribution optimizes the predictive accuracy and finds its hypothesis is wrong. Next,this paper proposes a method which employs a Cross Validation method to search an ESS value to optimize the predictive accuracy and shows the effectiveness of the proposed method.

1. はじめに

ベイジアン・ネットワークの学習は AI 分野内の基礎研究トピックの一つであり、AI の研究者や統計学者はこれまで以下のような様々な Scoring Metrics を提案してきた。

[Cooper and Herskovits,1991][Cooper and Herskovits,1992] は一様分布と一般的なディレクレ分布を想定し、UPSM(Uniform Prior Score Metric) と DPSM(Dirichlet Prior Score Metric) の二つの score metric を提案している。

[Buntine,1991] はディレクレ事前分布を想定し、尤度等価 (Likelihood Equivalence) を満たすハイパーパラメータを紹介した。

[Suzuki,1993] はハイパーパラメータとして 1/2 を持つ DPSM の近似である、新しい MDL(Minimum Description Length) code を提案し、ハイパーパラメータが 1/2 の時のみ、DPSM が MDL に収束することを証明した。

[Lam and Bacchus,1994] はベイジアン・ネットワークでエンコードした MDL を提案した。

[Bouckaert,1994-1][Bouckaert,1994-2] は UPSM(事前分布が一様分布を持つ時の DPSM、またはハイパーパラメータが 1.0 の時の DPSM) が MDL に収束することを証明した。

[Heckerman et.al.,1995] は尤度等価推定法を提案し、一定値のハイパーパラメータを持つ Dirichlet prior がその推定を満たすための十分条件であることを示し、UPSM が尤度等価推定を満たさないことを指摘し、新しい score metric を BDe(likelihood-equivalence Bayesian Dirichlet score metric) と呼んだ。また、[Buntine,1991] のハイパーパラメータを BDe の特別な場合であるとし、それをを BDeu(Bayesian Dirichlet equivalent uniform) metric と呼んだ。

[Suzuki,1998] は [Bouckaert,1994-1] の展開が間違っていること、ハイパーパラメータの値が 1/2 の時のみ DPSM が MDL に収束することを主張した。

[Sailander,et.al.,2007] は、BDeu score がハイパーパラメータの ESS(Equivalent Sample Size) に対して非常に敏感であることを示した。そのため、近年経験的 BDeu によるベイジアン・ネットワークの学習が行われるようになってきた。

[Ueno,2008] は DPSM と BDe が強一致性を持つことを示した。また、ESS の役割を分析し、最適な ESS は真の構造とデータに影響を受けることを示し、有効な学習法として経験ベイズ法を用いたベイジアンネットワークの学習手法を提案した。

[Steck,2008] はデータから直接 AIC を最小にする ESS の値を推定する手法を提案している。

これらの研究では, 暗黙に ESS を変化させ BDeu を最大にすることが、現実のベイジアン・ネットワークの予測精度を最適化すると仮定してきた。しかし、この仮定の妥当性についてはこれまで全く議論されてきていない。本研究では、ニュートン法を用いて BDeu score を最大化したときの予測精度を調べる。その結果、BDeu score を最大化する ESS の値が必ずしも予測精度を最適化していないことが分かった。

そこで経験的ベイジアン・アプローチの観点から、Cross-Validation により予測精度を最大化するための漸近的に最適な ESS の値を求める手法を提案する。Cross-Validation では training データ及び validation データから学習され推定された構造間のエラーを最小化する ESS の値を探索することで、予測精度を最大化するための漸近的に最適な ESS の値を探索することが可能となる。結果として、この手法が最も良い学習精度であることを示す。

2. ベイジアン・ネットワーク

ベイジアン・ネットワークモデルは、図 1 のように対象とする現象をグラフで表す。

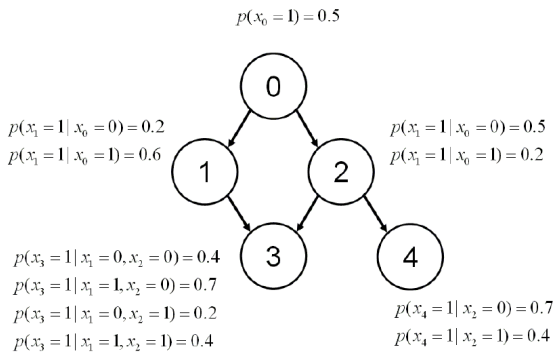

図 1: ベイジアン・ネットワーク

連絡先: 電気通信大学大学院, 〒 182-8585 東京都調布市調布ヶ丘 1-5-1

予測対象の各変数をノード、変数間の確率依存関係をアークとして確率ネットワークを構築したもので、確率構造を表すDAG(Directed Acyclic Graph)S と条件付き確率パラメータ集合 Θ_S で表される。

ベイジアン・ネットワークの同時確率分布は、条件付き確率パラメータよりチェーン・ルールを用いて以下のように表わすことができる。

$$p(x_1, x_2, \cdots, x_N) = \prod_{i=i}^{N} p(x_i | x_1, x_2, \cdots, x_{i-1}) \qquad (1)$$

ベイジアン・ネットワークにおいて、構造 S を所与とし、変数 i の親ノード集合を $\Pi_i \subseteq \{x_1, x_2, \cdots, x_N | S\}$ として、同時確率分布を以下の式で表わすことができる。

$$p(x_i, x_2, \cdots, x_N | S) = \prod_{i=1}^{N} p(x_i | \Pi_i, S) \qquad (2)$$

3. BDeu score

3.1 事後分布と予測分布

θ_{ijk} を親ノード変数集合 Π_i が j 番目のパターンとなった時の $x_i = k$ となる条件付確率を示すパラメータとする。このとき、データ $\mathbf{X}$ を得たときの尤度は以下の通りである。

$$p(\mathbf{X} | \Theta_S, S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\sum_{k=0}^{r_i-1} n_{ijk}!}{\prod_{k=0}^{r_i-1} n_{ijk}!} \prod_{k=0}^{r_i-1} \theta_{ijk}^{n_{ijk}} \qquad (3)$$

事前分布として以下のような共役自然事前分布であるディレクレ分布を想定することで、

$$p(\Theta_S | S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=0}^{r_i-1} \alpha_{ijk})}{\prod_{k=0}^{r_i-1} \Gamma(\alpha_{ijk})} \prod_{k=0}^{r_i-1} \theta_{ijk}^{\alpha_{ijk}-1} \qquad (4)$$

以下のような事後分布を得る。

$$\begin{aligned} p(X, \Theta_S | S) &= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=0}^{r_i-1} (\alpha_{ijk} + n_{ijk} - 1))}{\prod_{k=0}^{r_i-1} \Gamma(\alpha_{ijk} + n_{ijk} - 1)} \\ &\times \prod_{k=0}^{r_i-1} \theta_{ijk}^{\alpha_{ijk}+n_{ijk}-1} \end{aligned} \qquad (5)$$

ここで、MAP(maximum a posterior) 推定量は $\widehat{\theta_{ijk}} = \frac{\alpha_{ijk}+n_{ijk}}{\alpha_{ij}+n_{ij}}$、ただし、$n_{ij} = \sum_{k=0}^{r_i-1} n_{ijk}$、$\alpha_{ij} = \sum_{k=0}^{r_i-1} \alpha_{ijk}$ である。

さらにこれらより、以下の予測分布の式を得る。

$$\begin{aligned} p(\mathbf{X} | S) &= \int_{\Theta_S} p(\mathbf{X}, \Theta_S | S) p(\Theta_S) d\Theta_S \\ &= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \end{aligned} \qquad (6)$$

これは一般的に DPSM と呼ばれる。

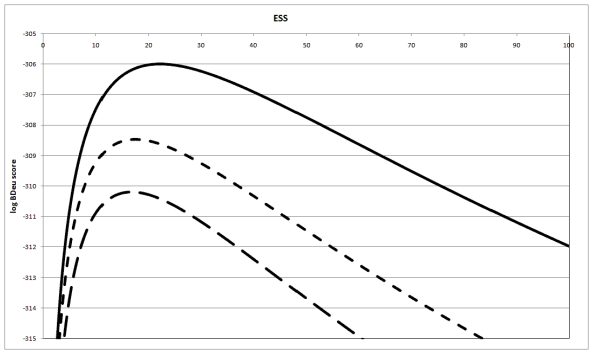

図 2: log BDeu score

3.2 BDeu score

BDeu score は式 (6) におけるハイパーパラメータの設定手法である。この手法を用いることで、ハイパーパラメータに事前知識を取り入れることができる。BDeu score では、ハイパーパラメータに一様な事前制約 $\alpha_{ijk} = \frac{\alpha}{r_i q_i}$ を使用する。ここで、α はユーザによって任意に決められた Equivalent Sample Size(ESS) である。よって BDeu score は以下の式から求めることができる。

$$p(\mathbf{X} | S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\frac{\alpha}{q_i})}{\Gamma(\frac{\alpha}{q_i} + n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(\frac{\alpha}{r_i q_i} + n_{ijk})}{\Gamma(\frac{\alpha}{r_i q_i})} \qquad (7)$$

BDeu score の特性として、図 2 のように ESS の値について単峰性を持つことが分かっている。

4. 予測分布の最大化

これまでの研究では、暗黙に ESS を変化させ BDeu を最大にすることが、現実のベイジアン・ネットワークの予測精度を最適化すると仮定してきた。しかし、この仮定の妥当性についてはこれまで全く議論されてきていない。本研究では、ニュートン法を用いて BDeu score を最大化したときの予測精度を調べる。

BDeu score が単峰性を持つことに着目し、ニュートン法を用いた経験ベイズ学習を提案する。

式 (7) において、ガンマ関数の特性 $\Gamma(z) = (z-1)\Gamma(z-1)$ より次の形にできる。

$p(\mathbf{X} | S) =$

$$\prod_{i=1}^{N} \prod_{j=1}^{q_i} \left[\prod_{t=1}^{n_{ij}} \frac{1}{\frac{\alpha}{q_i} + n_{ij} - t} \prod_{k=0}^{r_i-1} \prod_{t'=1}^{n_{ijk}} \left(\frac{\alpha}{r_i q_i} + n_{ijk} - t' \right) \right] \qquad (8)$$

よって log BDeu score は次のようになる。

$\log p(\mathbf{X} | S) =$

$$\sum_{i=1}^{N} \sum_{j=1}^{q_i} \left[\begin{array}{l} -\sum_{t=0}^{n_{ij}-1} \log \left(\frac{\alpha}{q_i} + t \right) \\ +\sum_{k=0}^{r_i-1} \sum_{t'=0}^{n_{ijk}-1} \log \left(\frac{\alpha}{r_i q_i} + t' \right) \end{array} \right] \qquad (9)$$

また、式 (9) の α についての一階偏微分の式 $f(\alpha)$ は次のよ

表 1: シミュレーション結果

構造	sample 数	Newton Method	BDeu(0.5)	BDeu(1.0)	Cross-Validation
Bayesian Network (1)	100	237	8	26	342(α=10)
	500	867	887	940	979(α=2)
	1000	933	1000	998	998($\alpha = 1$)
Bayesian Network (2)	100	101	62	77	443(α=6)
	500	325	97	155	771(α=6)
	1000	635	365	413	995(α=2)
Bayesian Network (3)	100	2	0	0	9(α=89)
	500	5	0	0	31(α=96)
	1000	27	0	0	59(α=100)

うに得られる。

$$f(\alpha) = \sum_{i=1}^{N}\sum_{j=1}^{q_i}\left[\begin{array}{l} -\sum_{t=0}^{n_{ij}-1}\frac{1}{q_i}\frac{1}{\frac{\alpha}{q_i}+t} \\ +\sum_{k=0}^{r_i-1}\sum_{t'=0}^{n_{ijk}-1}\frac{1}{r_i q_i}\frac{1}{\frac{\alpha}{r_i q_i}+t'} \end{array}\right] \quad (10)$$

BDeu score の単峰性により、式 (10) がにおいて $f(\alpha)$ =0 となる時の α の解が予測分布を最大化する ESS の値となる。

予測分布の最大化による予測精度の影響を確認するために、以下の手順で実験を行った。

1. 図 1 の構造 (条件付き確率を一様に分散させて設定した構造) Bayesian Network(1) よりサンプル数 100,500,1000 個のデータを生成する。
2. 各データ毎に真の構造を所与として、予測分布を最大化する ESS の値を求め、その値を用いて構造推定をする。
3. 手順 1,2 を 1000 回行い、真の構造を正しく推定できた回数を調べる。

また、同様の実験を因果が強い構造 (条件付き確率を 0 または 1 付近に設定した構造)Bayesian Network(2) と因果が弱い構造 (条件付き確率を 0.5 付近に設定した構造)Bayesian Network(3) について行った。さらに比較対象として、ESS の値を 0.5 と 1 に固定した BDeu score についての実験も行った。

実験の結果を表 1 にまとめる。表の”Newton Method”は Newton 法により予測分布を最大化する学習の結果、”BDeu(0.5)”と”BDeu(1)”はそれぞれ ESS の値を 0.5 と 1 に固定した BDeu score を用いて学習を行った結果である。

これまでのベイジアン・ネットワークの研究では、予測分布を最大化することが予測精度を最大化することになると考えられてきた。しかし ESS の値を 0.5 や 1.0 に固定して学習を行った時と比較した時に、Bayesian Network(1) の時のサンプル数 500、1000 の時に予測精度が減少した。その結果から予測分布を最大化することが必ずしも予測精度を最大化するとは限らないことが分かった。その理由として、データから得られた予測分布は、データの歪みもそのまま反映してしまうために、真の構造についての予測分布ではなく、そのデータを最も良く表現できる構造についての予測分布を作るためであると考えられる。

5. 提案手法

経験的ベイジアン・アプローチの観点から、Cross-Validation によって予測精度を最大化するための漸近的に最適な ESS の値を探索する手法を提案する。

n 個のデータ $\mathbf{X}$ が与えられた時の Cross-Validation のアルゴリズムは以下の通りである。

1. ESS α の範囲を $a \leq \alpha \leq b$ とする。
2. ESS の初期値を $\alpha = a$ とする。
3. n 個のデータ $\mathbf{X}$ における 50%を training データとしてランダムにサンプルする。残ったデータを validation データと呼ぶ。
4. training データから α を持つ BDeu を用いて構造を学習する。
5. validation データから α を持つ BDeu を用いて構造を学習する。
6. training データと validation データから学習された推定構造のエラーをカウントする。一つのエラーは training データから学習された構造より validation データから学習された構造が親ノードが欠損していたり、余分にある時にカウントされる。
7. $\alpha \leq b$ になるまで ESS の値を $\alpha = \alpha + 1$ とする。
8. 手順 2-7 を 10 回繰り返す。
9. 10 回の試行のエラーの平均を計算する。
10. エラーの平均を最小にするための ESS の値を選択する。

このアルゴリズムを用いることで、BDeu score を用いた学習の予測精度を最大化するための漸近的に最適な ESS の値を探索することができる。

提案した手法の有効性を調べるために以下の手順で実験を行った。データは 4.2 で発生させたデータを用いた。

1. 3 種類の構造の各サンプル数における ESS の値を上記のアルゴリズムを用いて決定する。
2. 手順 1 で求めた ESS の値を用いて 1000 セットのデータで学習を行い、真の構造を正しく推定できた回数を調べる。

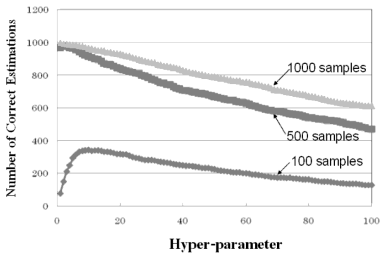

図 3: Bayesian Network(1)

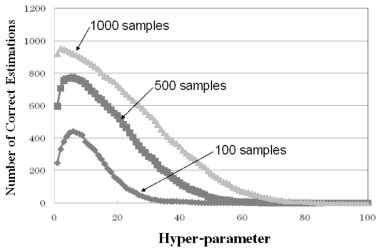

図 4: Bayesian Network(2)

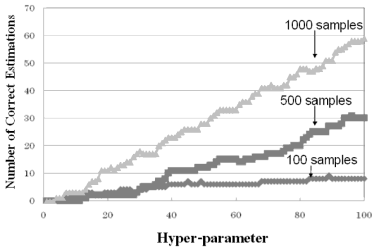

図 5: Bayesian Network(3)

実験の結果を表 1 の”Cross-Validation”に示す。

表 1 より、Cross-Validation を用いて予測精度を最大化する手法が最も精度が良いことが分かる。Cross-Validation を用いた手法の結果が良いのは、この手法で求める ESS の値は予測分布によらず、直接予測精度を最大化するような ESS の値を求めているために結果が良いと考える。

また、ESS の値を変化させた時の構造一致回数の様子を図 3、4、5 に示す。図 3 は Bayesian Network(1)、図 4 は Bayesian Network(2)、図 5 は Bayesian Network(3) の様子をそれぞれ表わす。

6. むすび

これまでの Bayesian Network の構造学習についての研究では、暗黙に ESS の値を変化させ BDeu を最大にすることが、現実のベイジアン・ネットワークの予測精度を最大化すると仮定していたが、この仮定の妥当性についての議論はなかった。よって本研究ではまず、ニュートン法を用いて予測分布を最大化する ESS の値を求め、その値を用いた時の学習精度を調査した。その結果として、予測分布を最大化することが、必ずしも予測精度を最大化しないことを示し、これまでの仮定が間違っていることが分かった。

次に、経験的ベイジアン・アプローチの観点から Cross-Validation により予測精度を最大化する ESS の値を求める手法を提案し、実験を行った。その結果、この手法によって求めた ESS の値を用いた学習が最も良い学習精度を示すことが分かった。

参考文献

[Bouckaert,1994-1] Bouckaert,R.: ”Probablistic network construction using the minimum description length principle”, Technical Report RUU-C9-94-27,Utrecht University.(1994)

[Bouckaert,1994-2] Bouckaert,R.: ”Properties of Bayesian network learning algorithm”, *Proc.Uncertainty in Artificial Intelligence,California*,102-109.(1994)

[Buntine,1991] Buntine,W.: ”Theory refinement on Bayesian networks”, *Proc.Uncertainty in Artificial Intelligence*,52-60.(1991)

[Cooper and Herskovits,1991] Cooper, G.F. and Herskovits, E.H.: ”A Bayesian Methods for the induction of probabilistic networks from data”, Technical Report SMI-91-1,Section on Medical Informatics,Stanford university.(1991)

[Cooper and Herskovits,1992] Cooper,G.F. and Herskovits,E.: ”A Bayesian Methods for the induction of probabilistic networks from data”, *Machine Learning*,9,309-347.(1992)

[Heckerman et.al.,1995] Heckerman,D.,Geiger,D., and Chickering,D.M.: ”Learning Bayesian networks:The combination of knowledge and statistical data”, *Machine learning*,**20(3)**,197-243.(1995)

[Lam and Bacchus,1994] Lam,W. and Bacchus,F.: ”Learning Bayesian Belief Networks:An Approach based on the MDL Principle”, *Computational Intelligence*,**10.4.**,269-293.(1994)

[Sailander,et.al.,2007] Silander,T.,Kontakanen,P.,& Myllymaki,P.: ”On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter”, *Proc.the twenty-third conference of Uncertainty in Artificial Intteligence*,pp.360-367,(2007)

[Steck,2008] Steck,H.:”Learning the Bayesian Network Structure:Dirichlet Prior versus Data”, *proc.Uncertainty in Artificial Intelligence*,pp.511-518(2008)

[Suzuki,1993] Suzuki,J.: ”A Construction of Bayesian networks from Databases on an MDL Principle”, *Proc.Uncertainty in Artificial Intelligence*,266-273.(1993)

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

[Ueno,2008] Ueno,M.:”Learning likelihood-equivalence Bayesian networks using an empirical Bayesian approach”, *Behaviormetrika*,Vol.35,No.2,pp.115-135,(2008)

[Yang and Chang,2002] 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)