

決定的従属関係の含まれるベイジアンネットワークの近似推論手法に関する考察

中島 祐輝[†]
Yuki Nakajima

戸田 貴久[‡]
Takahisa Toda

1 はじめに

ベイジアンネットワーク (BN) とは因果関係を有向非巡回グラフと確率で表したものであり、確率変数、有向辺、条件付き確率表から構成されている。BN を活用することで、確率推論を行うことができるが、厳密に推論結果を求めることは計算量等の課題がある。そこで、ギブスサンプリングや尤度重み付きサンプリングを活用することで近似的に推論を行うことができるが、決定的な従属関係の数やエビデンスによっては推論精度が悪化する可能性がある。

本研究では一様分布からサンプルを生成し、重要度重みを掛けることで元の確率分布に従うサンプルとする重点サンプリングに基づいたサンプリングによる近似推論手法を提案する。提案手法では、SAT に基づくサンプリングを活用することで従来のサンプリング手法の課題点を解決を目指す。そして、従来手法と提案手法で推論精度の比較実験を行い、各手法の精度を評価する。

2 ベイジアンネットワーク (BN) の近似推論

BN の確率推論は、ある情報を観測したときに他の確率変数についての発生確率を求めることができる。ここでは観測された情報をエビデンス E 、求めたい確率変数とその値をクエリ X と呼び、確率推論は事後確率 $P(X|E)$ を求めるということになる。厳密な確率推論値を求める手法として確率伝搬法などがあるが計算量などの課題が存在する。

そこで、ギブスサンプリングなどでサンプルを生成し、エビデンスに一致するサンプルと、エビデンス・クエリに一致するサンプルの個数の比率を計算することで近似的に確率推論を行うことができる。条件付き確率表において確率値が 0.0, 1.0 となっている個所は、そのノードと親ノードが決定的な従属関係であることを示している。もし BN に決定的な従属関係が多く現れるとマルコフ連鎖が既約性を持たず、ギブスサンプリングにより生成されるサンプルが偏る可能性がある [1]。また、尤度重み付きサンプリングはエビデンスの数や発生確率によって重みが小さくなってしまい、推論精度が悪化する可能性がある。

3 充足可能性問題 (SAT) とその応用

SAT とは与えられた命題論理式が真となる割り当て (解) が存在するか判定する問題である。また、SAT に基づ

いたサンプリング手法が研究されており、これは命題論理式の解を一様ランダムにサンプリングすることを目的としている。本論文ではこれを SAT サンプリングと呼ぶ。代表的な手法として、unigen [4] などがある。

現実の問題を SAT に変換して、SAT ソルバーで解を求める解法が広く活用されている。SAT に変換することを SAT 符号化といい、Chavira ら [3] は BN を重み付きの命題論理式に SAT 符号化し、確率推論を重み付きモデルカウント (Weighted model counting; WMC) に変換して厳密推論を行う手法を提案した。WMC とは、命題論理式に存在する解の重みの総和を求める問題である。しかし、解の総数はノード数などに応じて指数的に増加するため計算困難な場合がある。

そこで Chakraborty ら [2] は近似的に WMC を行う、アルゴリズム WeightMC を提案した。WeightMC のアイデアは、XOR 制約という命題変数や定数を排他的論理和 (XOR) で結んで得られる式を加えて命題論理式の解空間をランダムに縮小させる。この縮小した解空間をセルと呼ぶ。この時セルに存在する解の重みの合計が、どのランダムに得られるセルでも概ね均等になるように XOR 制約を加える。そこからセルに存在する解の重みの総和を求めてセルの総数を掛けることで、命題論理式に存在する解の重みの総和を近似的に求める。

4 提案手法

提案手法では、一様分布からサンプルを生成し、生成したサンプルに重みを付けることで元の確率分布に従うサンプルとする重点サンプリングに基づいたサンプリング方式を採用する。一様サンプリングの方法は、BN を SAT 符号化して得られた命題論理式に対して、SAT サンプリング unigen を活用する。これにより、ギブスサンプリングでは偏ったサンプリングとなる可能性のある決定的な従属関係を多く持つ BN に対しても、影響を受けずサンプルを生成できると考える。

提案手法では、同時確率分布 $p(z)$ からサンプリングする代わりに、より単純な一様分布 $q(z)$ からサンプリングをする。この時、生成したサンプルに対して重要度重みを付けることで本来の確率分布に従ったサンプルとする。サンプル $z^{(l)}$ の重要度重み r_l の計算は以下の式 (1) で行う。

$$r_l = p(z^{(l)})/q(z^{(l)}) \quad (1)$$

一様サンプリングであれば全てのサンプルの生成確率 $q(z^{(l)})$ は等しくなるため、式 (1) において考えるのは、 $p(z^{(l)})$ のみとなり、その値は生成されたサンプルをもと

[†] 電気通信大学大学院情報理工学研究所 情報・ネットワーク工学専攻

[‡] 電気通信大学大学院情報理工学研究所

表 1: ランダムに生成したエビデンスを使って
近似推論した時の RMSE

BN	各手法による近似推論の RMSE			
	ギブス サンプリング	尤度重み付き サンプリング	WeightMC	提案手法
BN50_d76_E10	0.156	0.013	–	0.177
BN50_d76_E50	0.175	0.004	–	0.002
BN50_d87_E10	0.458	0.021	–	0.058
BN50_d87_E50	0.055	0.012	0.443	0.014
BN70_d71_E10	0.305	0.012	–	0.224
BN70_d71_E50	0.414	0.016	–	0.176
BN70_d83_E10	0.317	0.005	–	0.048
BN70_d83_E50	0.317	0.013	1.338	0.013

に, BN の条件付き確率表から計算できる. 提案手法では一様にサンプリングをして得られたサンプルに対して, 目的の確率分布の確率値に比例する重要度重みを与える. この時, サンプルは一様に得られるため重要度重みの高いサンプルも得られる. そのため, 尤度重み付きサンプリングでの重みが小さいサンプルばかりが生成されるといった課題は回避できると考える.

5 評価実験

5.1 実験設定

決定的な従属関係を多く持つ BN を使って, ギブスサンプリング, 尤度重み付きサンプリング, WeightMC, 提案手法で近似推論を行う. タイムアウトは 300 秒に設定する. 生成するサンプル数は 3000 個とし, 10 回近似推論を行い推論精度を RMSE で評価する. エビデンスは全ノードの 10%, 50% からランダムに生成したものと, 葉ノードの 80% をエビデンスとして生成したものを扱う.

5.2 実験結果

表 1, 表 2 に実験結果を示す. 時間内に求められなかったものはハイフンで示す. ギブスサンプリングによる近似推論の結果は悪化する傾向が見られた. 決定的従属関係の多さが原因で既約性が満たされず, 偏ったサンプリングになってしまい, 推論結果が悪化したと考えられる.

尤度重み付きサンプリングでは, 表 2 の結果を見ると, 部分的に RMSE が非常に大きくなっており, 近似推論の精度が悪化していることが確認できる. そこで尤度重み付きサンプリングで生成したサンプルを確認すると, 生成した 3000 個のサンプルのうち, ほとんどが重みが 0 になっていることが確認できた. これはエビデンスが特定の厳しい条件下のみでしか発生しないエビデンスであったため, 重みが 0 となるサンプルばかりが生成されてしまい, 推論精度に悪影響を及ぼしたと考える.

WeightMC はほとんどがタイムアウトとなり, 計算が終わったものは誤差が大きい結果となった. タイムアウトの原因は, XOR 制約を加えすぎた結果, 解空間を縮小させすぎてしまい解が存在しない状態が続いていることが確認できた. また, 今回 SAT 符号化で得られた命題論理式の解の重みが, 大きい解と小さい解の差が非常に大きかった. そのため XOR 制約を加えてセルに存在する解の重みの合計がどのセルでも概ね均等にすることが困難であったため, 時間がかかる上に誤差も大きくなったと考える.

表 2: 葉ノードからランダムに生成したエビデンスを使って
近似推論した時の RMSE

BN	各手法による近似推論の RMSE			
	ギブス サンプリング	尤度重み付き サンプリング	WeightMC	提案手法
BN30_d73	0.199	0.011	–	0.030
BN30_d89	0.166	0.017	–	0.013
BN50_d76	0.236	0.019	–	0.039
BN50_d87	0.304	0.480	–	0.024
BN70_d71	0.369	0.016	–	0.193
BN70_d83	0.317	0.255	–	0.014

提案手法による近似推論の精度を評価する. 表 1, 表 2 の RMSE は比較的小さいが一部結果は大きい値になっている. この原因について, 解の総数に対してサンプル数が少なかったことが原因であると考えられる. 特にこの実験で使用した, BN70_d71_E10 は 6,946,816 個, BN70_d71_E50 の解の総数は 41,984 個も存在していることが確認できた. そのため, 一様サンプリングで 3000 個サンプルを生成するだけでは, 確率分布を近似するのに不十分だったため, 近似推論の精度が悪化したと考える.

6 まとめ

本研究では, 一様分布からサンプルを生成し, 生成したサンプルに重みを付けることで元の確率分布に従うサンプルとする重点サンプリングに基づいたサンプリング手法を提案した. ベースとなるサンプリングで一様サンプリングを活用することで, BN の構造やエビデンスによってギブスサンプリングや尤度重み付きサンプリングで生じる課題を解決できを確認できた. 一方でサンプルを生成する計算時間は既存手法に比べて大きいことが課題として残されている. また, BN が決定的な従属関係を多く持っていたり, エビデンスの数が多いと, ノードの値の取りうる組み合わせが減少するため, BN を SAT 符号化して得られる命題論理式の解の総数が減少することにより, 提案手法による近似推論の精度が良くなることが期待できると考える.

参考文献

- [1] Poon, Hoifung, and Pedro Domingos. "Sound and efficient inference with probabilistic and deterministic dependencies." *AAAI*. Vol. 6. 2006.
- [2] Chakraborty, Supratik, et al. "Distribution-aware sampling and weighted model counting for SAT." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 28. No. 1. 2014.
- [3] Chavira, Mark, and Adnan Darwiche. "On probabilistic inference by weighted model counting." *Artificial Intelligence* 172.6-7 (2008): 772-799.
- [4] Chakraborty, Supratik, et al. "Balancing scalability and uniformity in SAT witness generator." *2014 51st acm/edac/ieee design automation conference (dac)*. IEEE, 2014.