F-006

ベイジアンネットワークを用いた大規模 Web 推薦システムの開発

Development of Large Web Recommendation System Using Bayesian Networks

山崎 敬広† ソンムァン ポクポン† 石山 洸‡ 髙田 健一郎‡ 植野 真臣†
Takahiro Yamazaki Pokpong Songmuang Ko Ishiyama Kenichiro Takada Maomi Ueno

1．はじめに

近年，様々な情報から個々のユーザの嗜好を推測し，コンテンツを推薦する技術の研究が盛んに行われている．筆者らは，大規模 Web サイトにおいて，個々のユーザの閲覧履歴からサイト内ページを推薦するシステムを開発中である．推薦の方法として，サイト内の各ページの閲覧確率の依存関係をモデル化し，ユーザごとに推論を行い閲覧確率が高いページを推薦する方式を考えている．これを受けて，本システムでは，多数の確率変数間の複雑な依存関係を柔軟にモデル化できるベイジアンネットワーク[1]を用い，サイト内各ページの閲覧回数を変数として，各ページの閲覧確率の関係をデータから学習し，モデル構築を行おうとしている．本稿では，この推薦システムに向けて，どのようなベイジアンネットワークを構築するべきかについて，検討する．大規模 Web サイトでの推薦システムにおけるベイジアンネットワークでは，大規模なデータに対して学習・推論ともに高速に動作し，かつ，実データに対する予測精度が良いことが求められる．これらの要求を満たすベイジアンネットワークを求めるために，MWST[2]，PolytreeMWST，AIC[3]，MDL[4]の 4 つの学習アルゴリズムを用いてそれぞれベイジアンネットワークを実データから構築し，予測精度と計算時間について比較実験を行った．その結果，MWST，PolytreeMWST が優れていることがわかった．さらに，本論ではこの MWST を大規模なデータに対して有効に用いるために，アプリオリアルゴリズムを応用した高速化アルゴリズムを提案する．実際に，大規模な実データに対して，提案アルゴリズムによるベイジアンネットワーク構築を行い，予測精度と計算時間を求めたところ，高い予測精度と計算時間の短縮を確認し，提案アルゴリズムの有効性を示した．

2．ベイジアンネットワークの学習アルゴリズム

ベイジアンネットワーク[1]とは，予測対象の各変数をノード，変数間の確率依存関係を有向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)と条件付確率パラメータ集合で表現される．本節では，比較実験を行った 4 つのベイジアンネットワーク学習アルゴリズムについて説明する．

2.1 MWST，PolytreeMWST

MWST[2]は，

$$I(A,B)=\sum_{A}\sum_{B}p(A,B)\ln\frac{p(A,B)}{p(A)p(B)}$$

で表される相互情報量を基準として木を構築する手法である．相互情報量の大きい順にノード間に枝を加えていくことにより，木構造を作り上げる．PolytreeMWST は，MWST が取り得る構造を，複数個親を持つことができる木構造(Polytree)に拡張したものである．どちらも計算量は高々$O(N^2)$である．

2.2 AIC，MDL

AIC[3]は，モデル選択のための情報量基準であり，期待対数尤度からの近似アプローチで求められている．MDL[4]は，AIC と同様にモデル選択のための情報量基準であり，情報量理論からのアプローチで求められている．ネットワークの構造探索は貪欲アルゴリズムで行い，その計算量はノード数 N に対し $O(2^N)$となる．学習で得られる構造は一般的な構造となるが，AIC のほうが MDL よりアークの数が多い複雑な構造になる傾向を持つ．

3．学習アルゴリズムの比較実験

どの学習アルゴリズムが推薦システムに有効であるか調べるために，学習アルゴリズムの比較実験を行った．

3.1 実験手順

以下に実験の手順を示す．

① 推薦システムを用いる Web サイトの中から閲覧数の多いページ 100 個を選び，各学習アルゴリズムを用いてベイジアンネットワークを構築する
② 100 個のページのうち，5 個以上のページを閲覧しているユーザ 20 人のデータをテストデータとする
③ ユーザが閲覧している 5 個のページについて，そのうちの 0～4 ページを証拠として与えて確率推論を行い，推論された確率の高い上位 8 つのページを推薦するページとする
④ 証拠として与えた以外のユーザが閲覧したページが推薦された 8 つのページに含まれる割合を予測精度として求める
⑤ この予測精度を証拠の個数ごとに全証拠パターンについて計算し，これらの平均をとったものを最終的な予測精度とする

3.2 実験結果

実験結果として得られた予測精度評価を図 1 に示す．グラフ横軸がネットワークに与えた証拠数，縦軸が 20 人のユーザデータの予測精度平均値を表している．これより証拠数を増やしていくと徐々に予測精度が高くなっていくことがわかるが，すべての証拠数の場合において，MWST と PolytreeMWST が AIC と MDL よりも高い予測精度を示していることがわかる．また，計算時間については，構造推定，推論，どちらの計算時間についても MWST と PolytreeMWST のほうが短く，高速に動作していた．これらのことから，MWST，PolytreeMWST が本システムにおいて有効であるといえる．

† 電気通信大学
‡ （株）リクルート

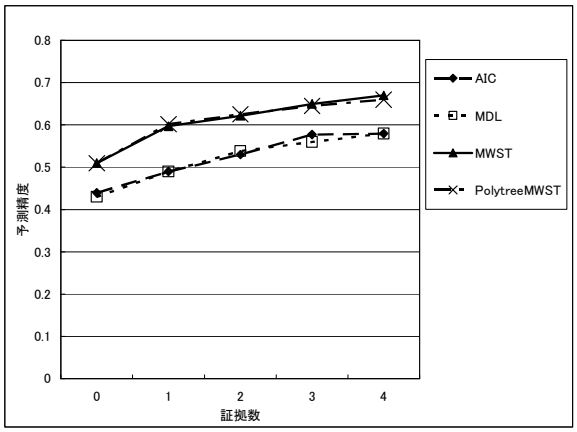

図 1. 予測精度の比較実験結果

4. 高速化アルゴリズム

比較実験を通して MWST，PolytreeMWST が予測精度と計算時間の両方について優れていることがわかった．しかしながら，高速であるこれらのアルゴリズムを用いても，ノード数が膨大になると計算時間が膨大になってしまう．これを解決する手法として，MWST の高速化アルゴリズムを提案する．

4.1 高速化アルゴリズムの理論

MWST では相互情報量が大きい順にノード間に枝を加えてベイジアンネットワークを構築する．このため，一般的にはすべてのノード間の相互情報量を求める必要があり，これに $O(N^2)$の計算量がかかるため，大規模データでは計算時間が膨大となる．そこで，高速化アルゴリズムでは，相互情報量が大きいアイテムの組み合わせを抜き出し，そのようなアイテムについてのみ相互情報量を計算することで，計算量の削減を行う．ここで，前提条件として，用いるデータが大規模なアイテム集合からの購買履歴データのような，あるアイテム I が選択される確率 P(I=1)が小さいデータとする．この場合，ある 2 つのアイテム A，B が同時に選択される確率 P(A=1, B=1)は 0 に近く，逆にどちらも選択されない確率 P(A=0, B=0)は 1 に近くなると考えられる．また，このことから P(A=1, B=0)~P(A=1)，P(A=0, B=1)~P(B=1)とできる．このような場合，P(A=1,B=0)，P(A=0,B=1)，P(A=0, B=0)が相互情報量に与える影響は非常に小さくなり，相互情報量はほぼ P(A=1, B=1)の値によってのみ決定される．このことから，相互情報量が大きなアイテムの組を探すことは，P(A=1, B=1)が大きなアイテムの組を探すことと同意になる．これにより，本来なら A,B のすべての確率を求める必要があったものを P(A=1,B=1)のみ求めればよくなるため，計算量を削減できる．さらに，P(A=1, B=1) = P(A=1)P(B=1|A=1)より，P(A=1, B=1)が大きい場合は，P(A=1)もしくは P(B=1)の値が大きくなるため，P(A=1)や P(B=1)が大きいアイテムを探索することも相互情報量が大きくなるノード組を探すことにつながるといえる．

4.2 高速化アルゴリズム

前節の理論を踏まえて，高速化アルゴリズムを以下に示す．

① P(X=1) > ε となるようなアイテム X をすべて選択する
② ①で選択されたすべてのアイテムについて P(X=1, Y=1) > ε となるようなアイテム X,Y をすべて選択する
③ ②で選択されたすべてのアイテムについて相互情報量を求め，MWST による構造推定を行う

4.3 高速化アルゴリズムの評価実験

推薦システムを用いる Web サイト内の 6042 個のページを持つサイトについて，高速化アルゴリズムを用いた構造学習を行い，100 人のユーザデータをテストデータとして，その予測精度を調べた．用いるページ，ユーザ数は違うが，それ以外は 3.1 節の実験手順と同じ手法で評価を行った．図 2 にその結果を示す．証拠を 4 つ与えた場合に予測精度が 0.7 を越えるなど，高い予測精度を示していることが分かる．また，このときの構造推定の計算時間は 33 分 27 秒であり，高速化アルゴリズムを用いない場合は 1 日以上の計算時間がかかっていたため，非常に高速化できたといえる．

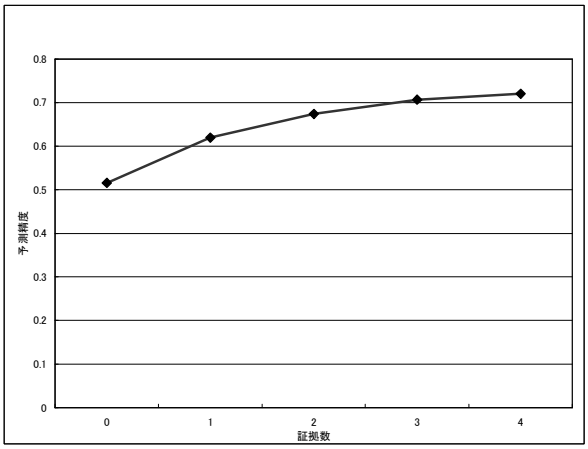

図 2. 高速化アルゴリズムによるネットワークの予測精度

5. まとめ

大規模 Web 推薦システムに向けて，予測精度の高いベイジアンネットワーク学習アルゴリズムを比較実験により決定した．さらに高速化アルゴリズムにより，そのアルゴリズムの高速化を行い，予測精度評価により有効性を示した．今後は，さらに色々なデータに対して実験を行い，評価を固める予定である．

参考文献

[1] 繁桝算男, 植野真臣, 本村陽一, “ベイジアン・ネットワーク概要”, 培風館, 2006.

[2] Chow, C. K. and Liu, C. N., “Approximating discrete probability distributions with dependence trees”, IEEE Transactions on Information Theory, IT-14, pp.462-467, 1968.

[3] Akaike, H., “A new look at the statistical model identification”, IEEE Transactions on Automatic Control, 19, pp.716-723, 1974.

[4] Rissanen, J., “A universal prior for integers and estimation by minimum description length”, Annals of Statistics, 11, pp.416-431, 1983.