2001 年 ベイジアンネットチュートリアル
Tutorial on Bayesian Networks (BN2001)
Tokyo, Japan, July 29, 2001.

ベイジアンネット構築システム BAYONET

Bayesian Netowork construction system: BAYONET

本村 陽一*
産業技術総合研究所 情報処理研究部門

Abstract: ベイジアンネットを各種の問題に適用するためには、適切なベイジアンネットモデルを構築する必要がある。これまでの多くのベイジアンネットソフトウェアでは主に確率推論部分の実装に注意が向けられており、利用者がベイジアンネットを構築することはそれほど容易であるとは言えなかった。そこで、ベイジアンネットを構築するために重要な、変数の決定、グラフ構造の選択、条件付き確率の獲得、の 3 つに対してこれを支援するためのシステムを開発したのでここに紹介する。システムは JAVA で記述され、JDBC により主要な各種データベースと接続して利用することができる。また、TCP/IP コネクションにより外部プログラムと接続して、ネットワークサーバーとして機能させることも考えられている。

1 はじめに

ベイジアンネットにより確率推論を行うためには，問題構造を良く反映した適切な変数，グラフ構造，条件付き確率を持ったモデルを構築する必要があるが，多くのベイジアンネットシステム (例えば Hugin[4]) ではこのモデルの設計をほとんど利用者に委ねているために，確率推論を行う以前のモデル化の段階で困難が生じることが多い．例えば，実際にベイジアンネットを構築しようとすると，注目する確率変数をどのように定義し，また変数間の主要な依存関係をどのように選ぶか，また条件付き確率を頻度データから求める時にデータ数が十分でない場合の問題をどう解消するか，などの問題がある．またユーザが構築したモデルが不適切なために，確率推論の精度が低下することもしばしば起こる．そこでベイジアンネットを実際に応用したいというニーズが増えるにしたがって，適切なモデル，確率推論の精度を高めるようなモデルを構築する作業を支援するシステムの必要性も高まっている．

本稿では適切なベイジアンネットモデルを構築するために，確率変数の決定，グラフ構造の決定，条件付き確率値の決定，の 3 つを支援するために開発したシステムについて紹介する．

*〒 305-8568 茨城県つくば市梅園 1-1-1 つくば中央第二，tel: 0298-61-5836，e-mail: y.motomura@aist.go.jp，URL: http://staff.aist.go.jp/y.motomura/

2 ベイジアンネット構築システム

いくつかの変数については値がわかっている時に、値のわかっていない変数がある特定の値をとる確率を計算したい。これによって、もっとも可能性の高い変数の値を推定したり、全ての可能性を考慮して各値がとる確率で平均した意志決定を行うことができる。この確率推論を実行するために、複数の確率変数の間の定性的な依存関係をグラフ構造によって表し，その間の定量的な関係を条件付き確率で表したモデルがベイジアンネットである．最初に簡単にベイジアンネットを概観する．

確率変数 X_i, X_j の間の条件付き依存性をベイジアンネットワークでは向きのついたリンクによって $X_i \rightarrow X_j$ と表し，X_i を親ノード，X_j は子ノードと呼ぶ．親ノードが複数あるとき子ノード X_j の親ノードの集合を $\pi(X_j) = \{X_1, \cdots, X_i\}$ と書くことにする．

この時の変数 X_j の値が親ノードの変数の値によって影響をうけるが、それが非決定的、つまり親ノードの値だけにはよらない不確実性がある時、この関係を子ノードの変数 X_j について親ノードの値を条件とする条件付き確率,

$$P(X_j \mid \pi(X_j)) \tag{1}$$

で表すことができる．この一つの子ノードについての関係はベイジアンネットの中で X_j を子ノード, $\pi(X_j)$ を親ノード群とする局所的な木になっている．

確率変数が離散的な場合，条件付き確率は全ての状態における確率値を並べた表，CPT(Conditional Probability Table) によって過不足なく表すことができる．例えば

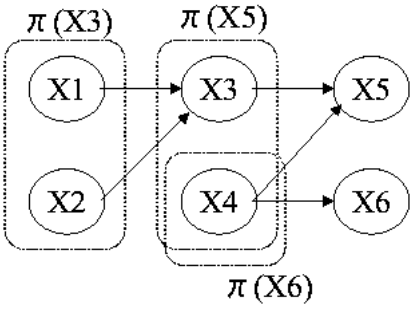

図 1: ベイジアンネットワークの例

親ノードがある状態 $\pi(X_j) = \boldsymbol{y}$ ($\boldsymbol{y}$ は親ノード群の各値で構成したベクトル) のもとでの n 通りの離散状態を持つ変数 X_j の条件付き確率分布を $p(X_j = x_1|\boldsymbol{y}), \cdots, p(X_j = x_n|\boldsymbol{y})$ とする (ただし $\sum_{i=1}^{n} p(x_i|\boldsymbol{y}) = 1.0$). これを行として，親ノードがとりえる全ての可能な状態 $\pi(X_j) = \boldsymbol{y_1}, \cdots, \boldsymbol{y_m}$ について列を構成した表 1 が X_j にとっての CPT，$P(X_j|\pi(X_j))$ である．

表 1: 条件付き確率表 (CPT)

$p(X_j = x_1\|\pi(X_j) = \boldsymbol{y_1})$	$\cdots$	$p(X_j = x_n\|\pi(X_j) = \boldsymbol{y_1})$
:	$\ddots$	:
$p(X_j = x_1\|\pi(X_j) = \boldsymbol{y_m})$	$\cdots$	$p(X_j = x_n\|\pi(X_j) = \boldsymbol{y_m})$

さて，ベイジアンネットを実際に構築する手順は以下の通りになる．

- モデルで使用する確率変数, X_j を決定しノードを作成する．
- 変数間の依存関係にしたがって，親ノードから子ノードにリンクを張っていく ($\pi(X_j)$ の決定).
- 変数間の依存関係を定量的に表す条件付き確率表 (CPT, $P(X_j|\pi(X_j))$) を決定する．

この 3 つを行う上で，データに照らして適切な判断を行うためにベイジアンネット構築システム[1]を開発した．本システムでは，データベースに格納された統計データを検索し，モデルとの適合性を確認しながら変数，局所木構造，条件付き確率というモデルの各要素を決定していく．モデル選択の基準となるのは尤度，情報量基準 MDL, AIC であり，これは利用者が自由に選ぶことができる．

システムは JAVA 言語により実装されており，豊富な GUI(グラフィックインターフェース) によるモデルの可視化，広く一般に使われている主要な各種データベースとの接続性の良さ，オブジェクト指向アーキテクチャによる拡張性の良さ，などを特色としている．

2.1 確率変数の決定を支援するデータベース操作機能

とくにこのシステム独自の特徴として，一般によく使われる SQL データベースと連携することで，従来はメモリ消費が激しく実行が難しくなるような大量のデータに対しても SQL 検索コマンドを用いた高度な操作が可能である．また，データとしては陽に格納されていないような変数 (例えばある 2 つの時刻の差分など) でも，SQL データベースの演算操作によって実現できればベイジアンネットの変数としてその場で利用することができる．具体的には，SQL データベース内の項目をベイジアンネットの変数に動的に割り当て，select 文で検索された該当データの頻度計算の結果を条件付き確率として取り込む．例えばスケジュールに関するデータベース中に Person という項目がありこれをノード X_{person} に割り当てると，その値 $\{motomura, \cdots\}$ を各状態とする確率変数を生成する．また come-in, go-out という入退出時刻を表す項目があったとすると，select 文の中で (go-out - come-in) のようにすることで両者の差分の滞在時間を求めることができる．これを離散化した $H = \{0, \cdots, 24\}$ をノード H に割り当て，$X_{person} \rightarrow H$ というグラフ構造を作ってデータをシステムに読み込むと，データベース中の各組合せの頻度をカウントして，さらにそれを正規化した条件付き確率 $P(H|X_{person})$ が得られる．これはある人が何時間滞在するかを表すモデルになっている．こうして得られた条件付き確率表のエントロピー[2]を計算し，(相互情報量の意味で) 条件付き依存関係の強さを評価することもできる．この例の場合，人によって滞在時間のばらつきが大きく，P_{person} を知ることで H の予測精度が高くなれば X_{person} と H の条件付き依存関係が強いことになる．このようにして適切な変数，リンクを決定していくことで，適切なベイジアンネットを構築していく．

2.2 グラフ構造の決定を支援する局所木選択機能

ベイジアンネットの場合，利用者が変数間にリンクを張って構築したグラフ構造が必ずしも適切であるとは限らない．そこで，候補となる複数のモデルの中からデータに対する尤度や情報量基準によって自動的に比較，選

[1] http://staff.aist.go.jp/y.motomura/

[2] 平均対数尤度に相当

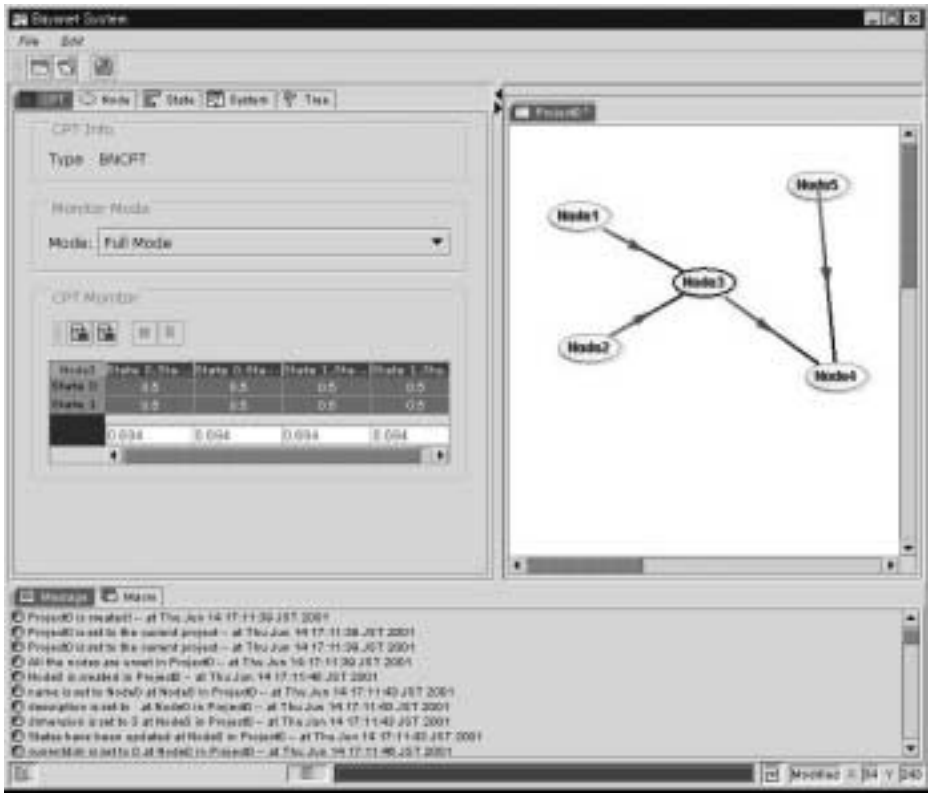

図 2: ベイジアンネット構築システム

図 3: データベース操作画面

択することが望まれている．この時，ベイジアンネットのグラフ構造は問題領域の因果構造を反映しており，モデルとしては単に尤度が高いだけではなく，グラフ構造がデータの発生過程 (因果構造) をできるだけ忠実に表していることが重要である．しかし変数が多くなるとありえるグラフ構造は組合せ的に増加し，探索空間が広大になるため，統計データだけから最適な予測モデルを得ることはそれほど簡単ではない．またデータだけからでは因果関係の順序 (リンクの向き) を読み取ることも難しい．そこでとくにグラフ構造についてはユーザの主観的判断も反映させながら，仮説として候補を限定し，様々な条件のもとでデータフィッティングを行いながら適切なものを決定することが必要となる．

子ノード 1 つを根，これに接続する親ノード群を葉とした木に注目すると，ベイジアンネットはこの木が組み合わさったものになっている．そして条件付き確率分布はこの局所木について一つ定義される．ここでベイジアンネットのグラフ構造の決定は各子ノード毎に最適な局所木を探索する Greedy アルゴリズムとして実現できる．つまり，(a) 子ノードを定義, (b) 子ノード毎に候補となる局所木を与える, (c) 各局所木ごとに条件付き確率を決定, (d) 最適な局所木を子ノード毎に Greedy に探索していく, という手順でベイジアンネットを構築する．(d) の手続きにおいて，木を選択する際に平均対数尤度とモデルの複雑性 (この場合は親ノード群の数) を考慮した選択基準 (MDL, AIC) によって事前に与えた候補集合の中から最適なものを選ぶ．

2.3 条件付き確率の決定を支援する学習機能

変数が離散的でかつ全ての組合せを含んでいる完全データの場合には，先に述べたような方法でデータベース中の頻度を計算し，最尤推定量を用いて CPT の全ての項を埋めることができる．

図 4: ノード定義画面

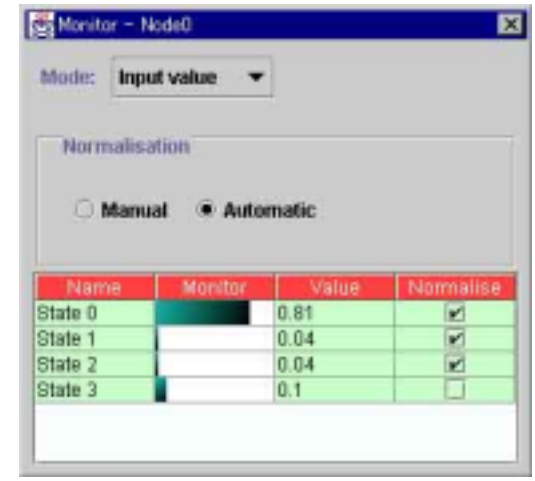

図 5: 確率値入力画面

一方，データが不完全で，変数の全ての組合せについてのデータが揃わないため，頻度から条件付き確率が得られないことがある．この場合は周辺のデータによって欠足している組合せの確率を推定することが必要となる．この未観測データの推定には親ノードの値を入力，その時の子ノードの確率値を出力とするようなニューラルネットや回帰モデルなどの学習モデルを利用することを考える．すでにわかっているデータの頻度から求めた確率値を教師信号として学習モデルで補間を行い，学習後にデータが欠足している親ノードの値に対するモデルの出力をその時の子ノードの確率，すなわち条件付き確率として用いる．これは学習モデルの汎化能力を期待して未学習の条件付き確率を近似していることになる．なおニューラルネット以外の他のパラメトリックモデルを利用するための拡張方法も用意されている．

図 6: グラフ構造の選択

図 7: 構造決定に用いる情報量基準の選択画面

図 8: 条件付き確率表 (CPT)

3　システムの拡張性，接続性

本システムは次のような特長により他のプログラムと連携して利用することができる．

- JDBC ドライバを持つ主要な各種データベースシステムとの接続
- JAVA のリフレクションによる，各種プログラムモジュールの追加
- 確率推論エンジン Hugin[4] との互換ファイル生成機能
- TCP/IP コネクションによる外部プログラムとのインタフェース

これらの特長により様々な問題に対する実用的な大規模なデータベースとの接続が容易になり，また条件付き確率の近似のための学習モデルとして，各種の回帰モ

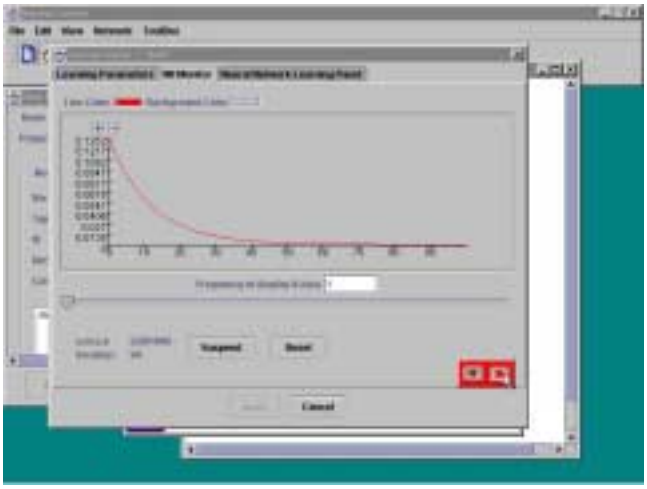

図 9: ニューラルネットによる条件付き確率の学習

デル，ニューラルネット，Support vector (regression) machine などのモジュールを容易に追加し，構造探索アルゴリズムとしては各種の情報量基準を追加，選択して性能評価を行うことが可能になる．本システムで構築したベイジアンネットは高速な Junction Tree Algorithm で定評のある確率推論エンジン，Hugin[4] でそのまま利用し，確率推論を実行することができる．本システムをフリーソフトとして公開しインプリメントに比較的手間のかかるグラフィカルユーザインターフェースやデータ管理部などを統一的に提供することで，各利用者は様々な最新の学習モデルやアルゴリズムの部分だけを実装し，短期間のうちに大規模データベースを利用して評価できるようになる．

4　まとめ

不確実性を含む問題領域において確率モデルの応用を進めるために，データベースと連携したベイジアンネットシステムを開発した．とくにこのシステム独自の特徴として，一般によく使われる SQL データベースと連携することで大量のデータに対して SQL 検索コマンドを用いた高度な操作を行い，適切な変数選択を支援する．また局所的に複数の木を作成しておき，その中から情報量基準にしたがって最適な木を自動的に選択することでグラフ構造を決定していく仕組みを導入した．さらにニューラルネットや回帰モデルなどを用いて学習することで欠足データがある場合やデータ数が十分でない場合でも条件付き確率値を近似することを可能にした．これらの特長によって，ユーザが問題構造に即した適切なベイジアンネットモデルを構築することが容易になる．本システムでは，局所木ごとのモデル，構造選択アルゴリズムなどを交換，拡張可能にしており，利用者が新たなモデル，アルゴリズムを追加することも容易にしている．これによって新たな理論的手法を実用的な規模の統計データにより短期間で評価することも可能になる．またこのソフトウェアを WWW や CD-R にて一般に無料で公開することで多くのユーザにより様々な問題に応用されることを期待している．

参考文献

[1] Castillo, E., Gutierrez, J., and Hadi, A.: *Expert Systems and Probabilistic Network Models*, Springer-Verlag (1997).

[2] 石塚満 (訳)：15 章: 確率的推論システム, 古川康一監訳，エージェントアプローチ人工知能, pp. 439–473 (1997), 共立出版.

[3] 本村陽一, 佐藤泰介, ベイジアンネットワーク–不確定性のモデリング技術–, 人工知能学会論文誌,*vol.15, No.4*, pp. 575–582 (2000).

[4] Hugin expert, http://www.hugin.com (2001).