BAYONET 新バージョンのご紹介

～ベイジアンネットモデリングのコツを交えて～

(株)数理システム 石田和宏

1. はじめに

ベイジアンネットワーク（以降ではベイジアンネットと略します）は様々な事象間の因果関係（厳密には確率的な依存関係）をグラフ構造で表現するモデリング手法の一つで、故障診断、気象予測、医療的意思決定支援、マーケティング、推薦システムなど様々な分野で利用や研究が行われています。

BAYONET はベイジアンネットモデリングのためのソフトウェアで、（独）産業技術総合研究所で開発され、数理システムがカスタマイズや機能追加を行い販売しています。今回の発表では本年リリース予定の新バージョンの紹介とベイジアンネットモデリングのちょっとしたコツをご紹介したいと思います。

2. ベイジアンネットについて

ベイジアンネットでは事象をノードで表現します。事象間に直接の確率的な依存関係があれば対応するノードを矢印で結びます。またその依存関係は条件付き確率表で定量的に表現します。グラフ構造は非循環有向グラフでなければなりません。

図１は病気の原因となる喫煙の有無や、病気の症状である呼吸困難といった観測できる情報から、観測できない肺がんなどの病気を診断するためのベイジアンネットです。

ベイジアンネットには次のような特徴があります。

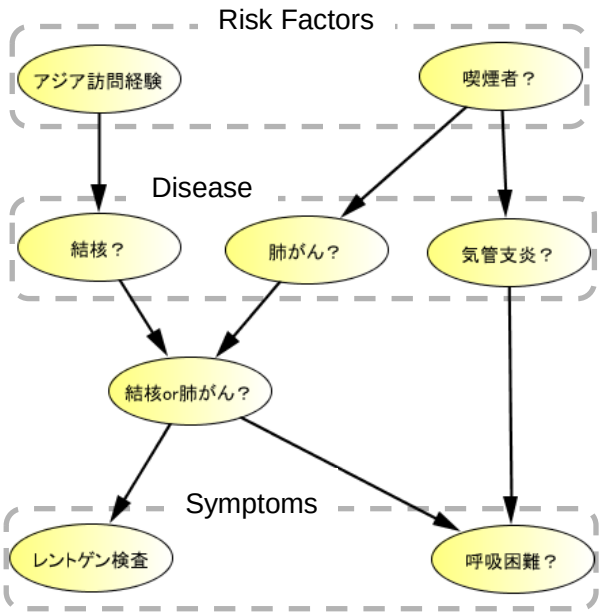


図 1　病気診断用のベイジアンネット

モデルの意味が理解できる

ベイジアンネットはグラフィカルモデルです。変数間の因果関係・依存関係をそのままモデルとして表現でき、視覚的にわかりやすいという特徴があります。

図１のベイジアンネットは病気の「リスクファクター」→「病気」→「症状」という三層の因果関係を表現しています。

予測を行うとき、説明変数の入力に欠損があってもよい

ベイジアンネットは予測時に全ての説明変数に値を入力する必要はありません。入力のない変数については条件付き確率表を元に適切にその影響を取り込みます。

モデルの利用用途が限定されない

ベイジアンネットは矢印の順方向だけでなく、逆方向にも推論が行えます。よって観測を入力する変数と予測対象となる変数がモデルで限定されることなく、自由に選択できます。

図１のベイジアンネットは病気を予測することが目的ですが、逆に肺がんを入力とすることにより、肺がんのリスクファクターや肺がんの症状を予測する目的でも利用できます。

3. BAYONET の新バージョンのご紹介

BAYONET は（独）産業技術総合研究所で開発され、数理システムが機能追加やカスタマイズを行い販売しているベイジアンネットモデリングのためのソフトウェアです。

昨年は 64bit 版をリリースしました。本年はモデル構築ウィザードをリニューアルします。現在の「モデル構築ウィザード」を「データのインポート」と「構造学習」の二つのウィザードに機能を分割します。

データインポートには簡単なデータの前処理機能を追加します。カラム名の変更、複数の値を一つの値にまとめるグルーピングや、連続値の離散化ができるようになります。

また構造学習では親子関係に制約を付ける GUI の一覧性が向上し、操作性を大きく改善します。また構造学習終了からモデル表示までのレスポンスを大幅に改善し構造学習全体にかかる時間を短縮します。

このほかにもさまざまな操作性の改善を行っています。

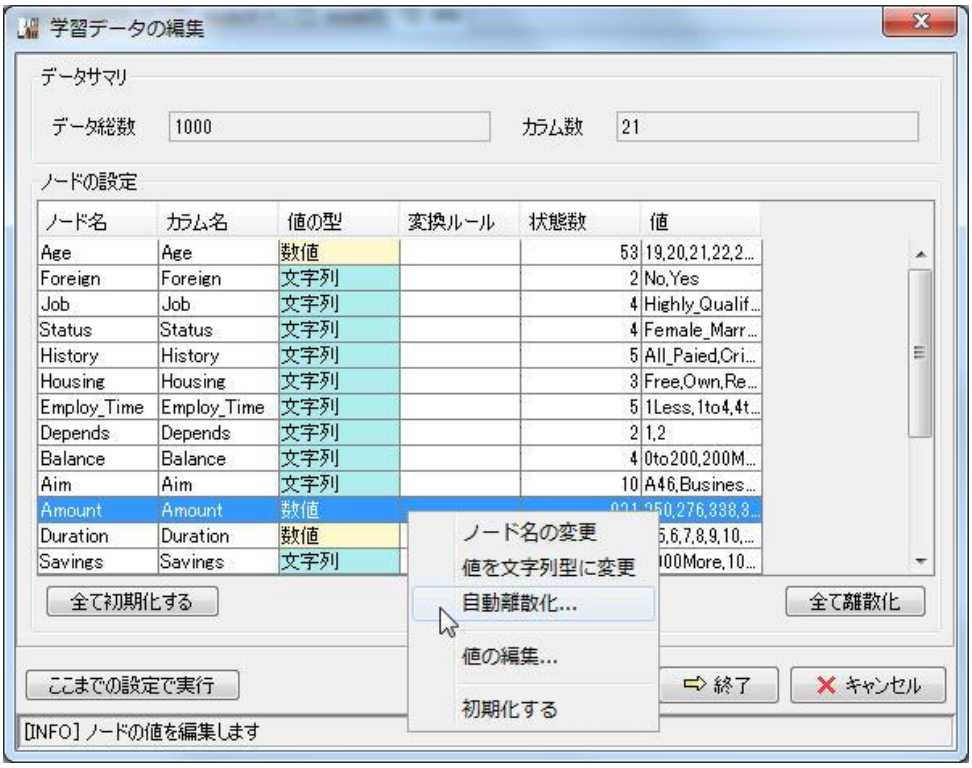


図 2　データインポートウィザード（前処理ページ）

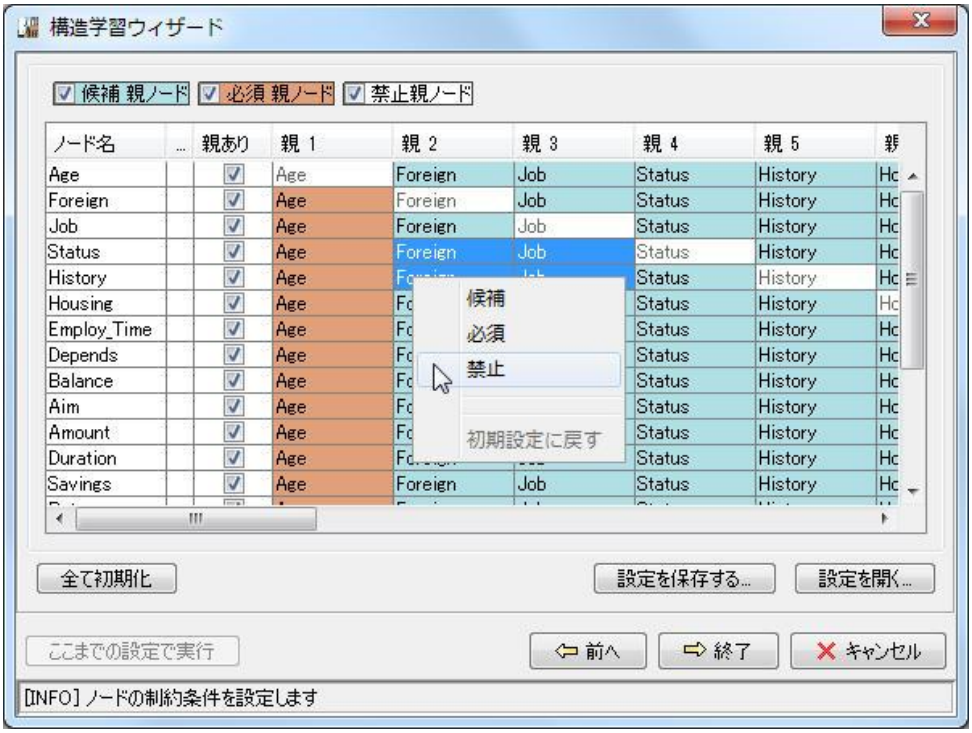
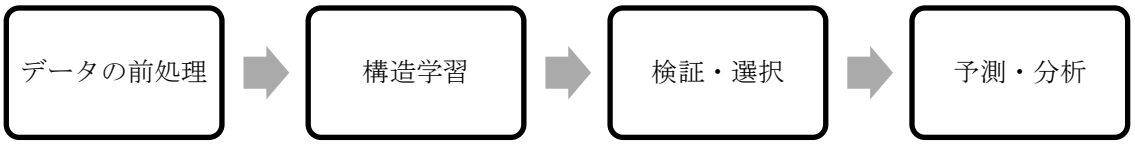


図 3　構造学習ウィザード（親子関係設定ページ）

4. ベイジアンネットモデル構築のプロセス

ベイジアンネットモデリングは一般的に次のようなプロセスとなります。BAYONET はこの各プロセスを支援するツールを提供します。



データの前処理（データインポートウィザード new）

BAYONET は数値データを直接扱えないため、連続変数は適切に離散化を行う必要があります。またベイジアンネットの要である条件付き確率表は子ノード・親ノード間のクロス集計表を正規化して作ります。このクロス集計表の各セルに、ある程度の頻度を持つようにすることが良いモデルを作る上で重要です。よってデータが少ない場合には、変数の状態数を最小限に抑えることが良いモデルを作るポイントの一つとなります。

ポイント

- 状態数が多い場合には適切に状態値をグルーピングします。これによりモデルの性能の向上が期待できます。
- 極端に頻度の少ない状態値は、欠損とするか他の状態値とマージするなどの処理が必要です。
- 「回答無し」のような状態値をそのまま使うと、回答の有無を判断するモデルになってしまう場合があり、適切に欠損とする必要があります。

構造学習（構造学習ウィザード new）

構造学習とは学習データから機械学習のアルゴリズムによりベイジアンネットのグラフ構造を決めることを言います。

ベイジアンネットのグラフ構造は基本的には変数間の因果の向きに合わせます（目的によってはその限りではありません）。しかしながらデータから因果の向きを決めることは難しい問題ですので、ユーザーが因果関係についていくつかの仮説を立て、実際にモデル構築を行い、その中から目的にあったモデルを選択するという手順になります。

数十変数を超える規模になってくると、その一つ一つに親子関係の仮説を立てるのは大変です。まずは変数を複数のグループに分類し、グループ間の関係の仮説を立てます。そしてグループ内での関係を検討し、これをあわせることで全体の仮説を作成します。

例えば図１のベイジアンネットであれば「Risk Factors」、「Disease」、「Symptoms」の３グループとなります。

仮説ができれば構造学習ウィザードでベイジアンネットを構築します。とりあえず変数間の関係を見てみたい場合は、親子関係については何も指定しないで構造学習を実行することもできます。この場合はすべての変数を親候補として構造を探索します。

ポイント

- 変数名にはグループの頭文字をプレフィクスとすると便利です。
- あまり細かく親子関係を規定すると、モデルの性能が悪化します。
- 親子関係は先入観にとらわれずいろいろな可能性を考えましょう。

モデル検証・選択（モデル検証ツール）

検証ツールを使って正答率などを参考にして目的にあったモデルを選択します。通常の予測モデルであれば説明変数と目的変数は一組しかありませんが、ベイジアンネットではどこを入力（説明変数）にして、どこを出力（目的変数）にしても良いという特徴があり、モデルの目的が一つとは限りません。この入力と出力の組を推論シナリオと呼びます。目的に応じて推論シナリオを複数考え検証を行います。

ポイント

- 推論シナリオは一つとは限りません。
- 推論シナリオが複数ある場合は、そのバランスを考えモデルを選択します。

予測・分析（推論ツール：エクセルアドイン）

予測・分析には推論ツール（エクセルアドイン）を使います。

分析では、説明変数への入力の組み合わせにより目的変数の確率分布がどのように変化するのかを確認します。説明変数が多い場合には、逆に目的変数に入力を設定し、説明変数の分布の変化を見ることにより有効な説明変数を絞り込むことができます。

影響の大きさを見る場合には事前分布との比もしくは差を見ることが多いです。またエクセルのグラフや書式設定を使うことで、視覚的にも分かりやすく表現することができます。

図 4　推論ツールの利用イメージ（1）

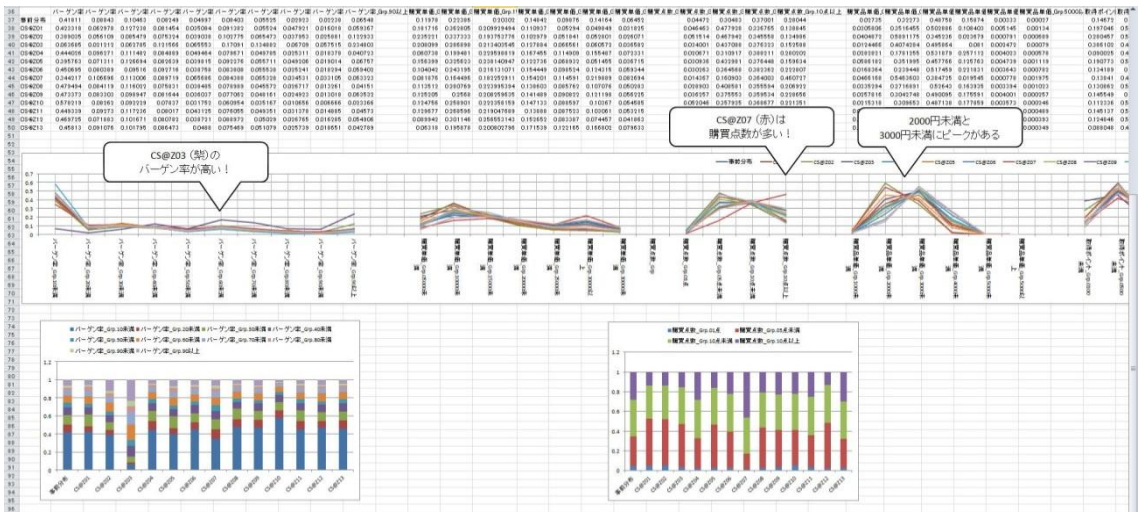


図 5　推論ツール利用イメージ（2）

ポイント

- 逆向き推論（目的変数に入力を入れる）によって、目的変数に影響を与える変数にあたりをつけることができます。
- 影響の大きさを見るには事前分布からの差・比をみます。
- エクセルの演算、グラフ、書式設定などの機能により分析の自由度が広がります。