BAYONET 新バージョンのご紹介

～Visual Mining Studio との連係、その魅力の機能～

株式会社 ＮＴＴデータ数理システム 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、NTT データ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。2002 年に販売を開始し、現在 300 以上のサイトで導入されています。昨年のバージョンアップでは「ベイジアンネットをより身近に」をコンセプトにユーザービリティを大幅に改善しました。さらに本年 9 月のバージョンアップでは、汎用データマイニングシステム Visual Mining Studio との連係により、充実の分析環境を実現しました。

Visual Mining Studio の持つ豊富なデータの加工や可視化、集計、分析などの機能が、ベイジアンネットモデリングの事前分析、学習データの加工、推論結果の可視化などに使えるようになりました。また Visual Mining Studio には最新の機械学習アルゴリズムが実装されており、それらとベイジアンネットとの比較も簡単に行えます。さらに Visual Mining Studio はビジュアルプログラミング環境を提供していますので、分析がマウス操作で簡単に行えます。分析の試行錯誤の過程が分析フローとして可視化、保存、だれでもアイコンダブルクリックでできるようになるのも大きな魅力です。

本発表ではベイジアンネットについて概要を説明した後、BAYONET の機能、新機能である Visual Mining Studio との連係についてデモを交えながらご紹介します。

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の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。

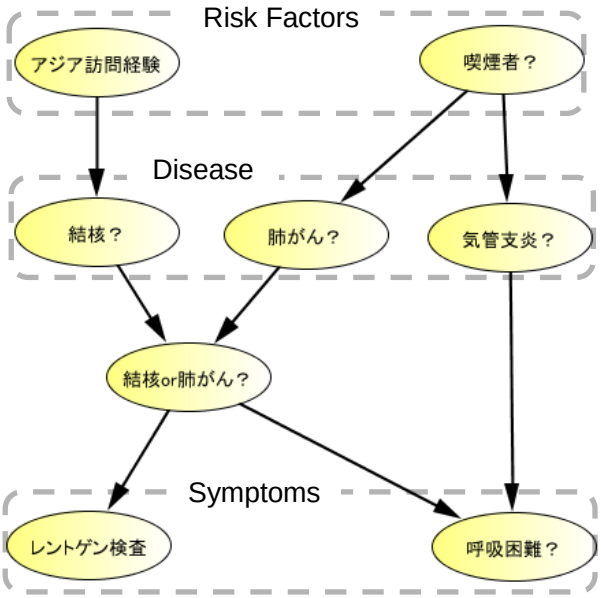


図 1　病気診断用のベイジアンネット

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。

予測を行うときに、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。

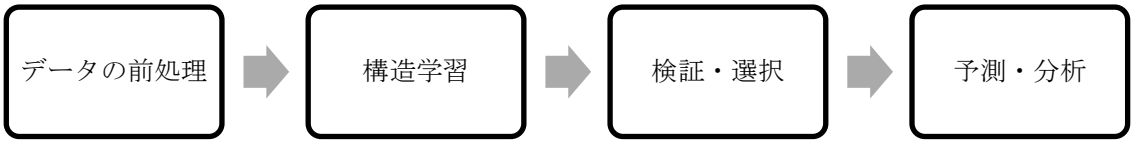
モデルの用途が限定されない

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

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す。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ています。



データの前処理（学習データのインポートウィザード）

BAYONET では構造学習を行う前に学習データのインポートを行います。この作業には「学習データのインポート」ウィザードを利用します。また、後述のようにデータマイニング Visual Mining Studio のデータをインプットとすることも可能です。

BAYONET では数値データを直接扱えません。数値データは適切に離散化を行い、カテゴリデータに変換する必要があります。またベイジアンネットのパラメータ、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計算します。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ためにグルーピングを行います。例えば 7 段階評価のアンケートであれば 3 段階にまとめるような操作を行います。

このような作業が「学習データのインポート」ウィザードによりマウス操作で簡単に行えます。

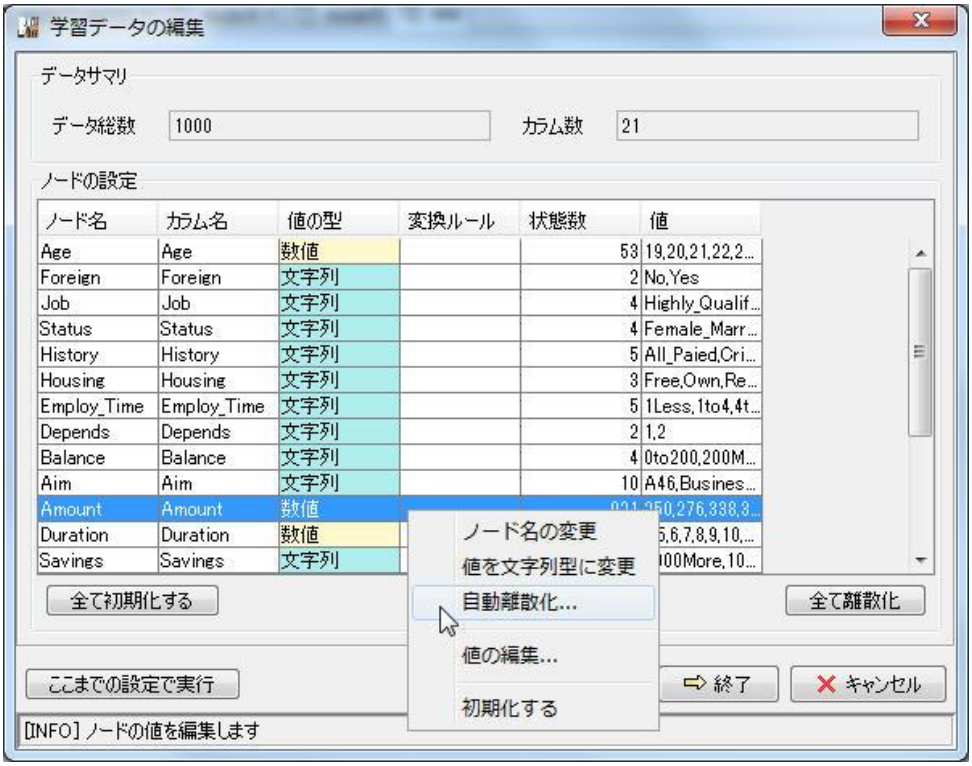


図 2　学習データのインポートウィザード

構造学習（構造学習ウィザード）

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。

構造学習ウィザードではこの因果関係の仮説の入力、学習アルゴリズム、評価基準の選択などを行います。

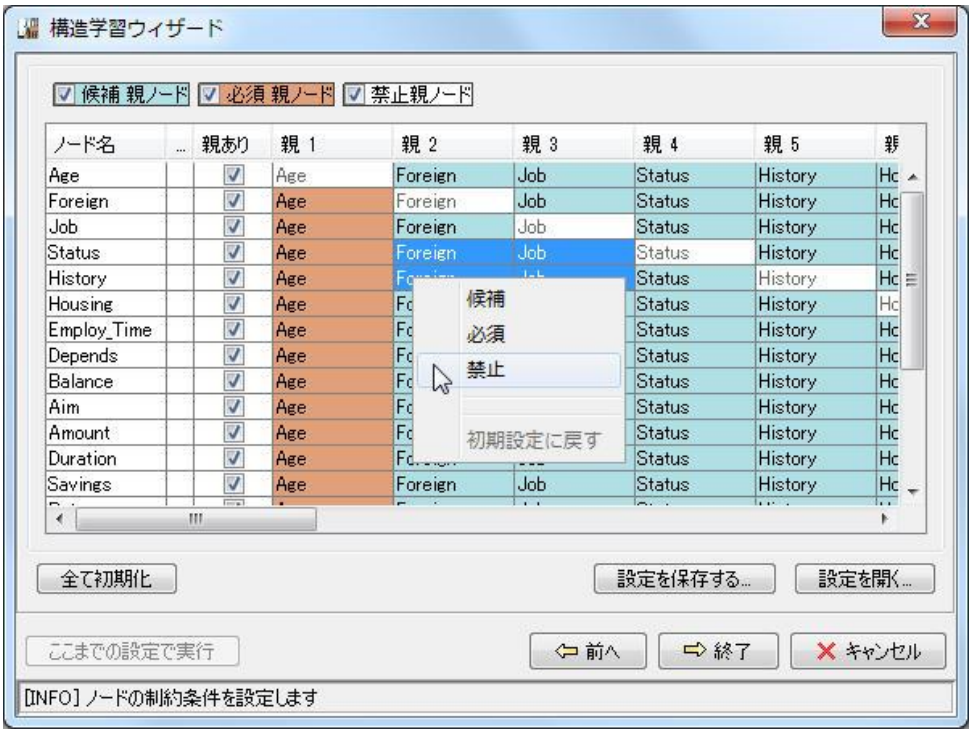


図 3　構造学習ウィザード

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計算し、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、選択を行います。

予測・分析（推論ツール：Excel 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計算します。また Excel 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。

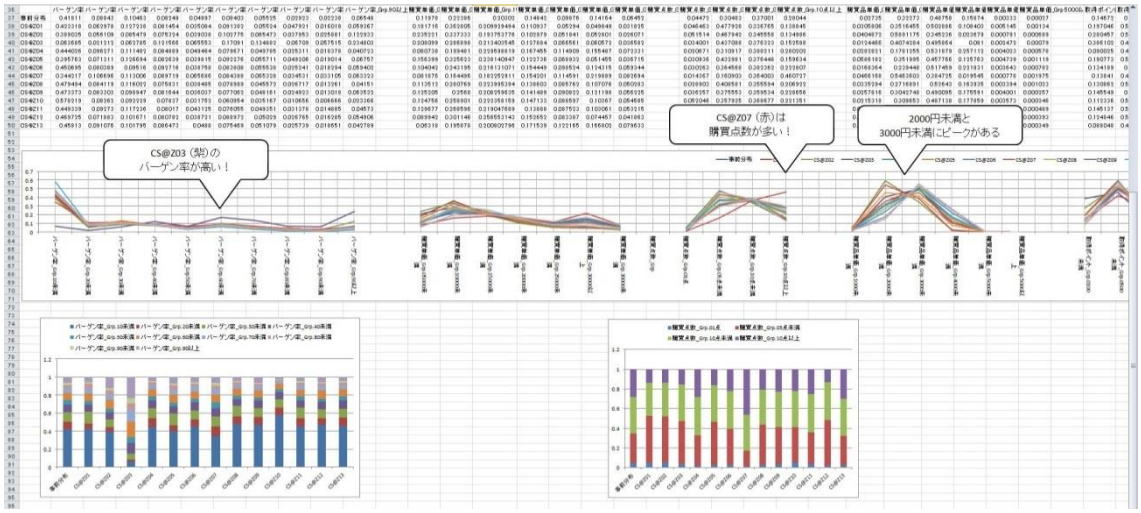


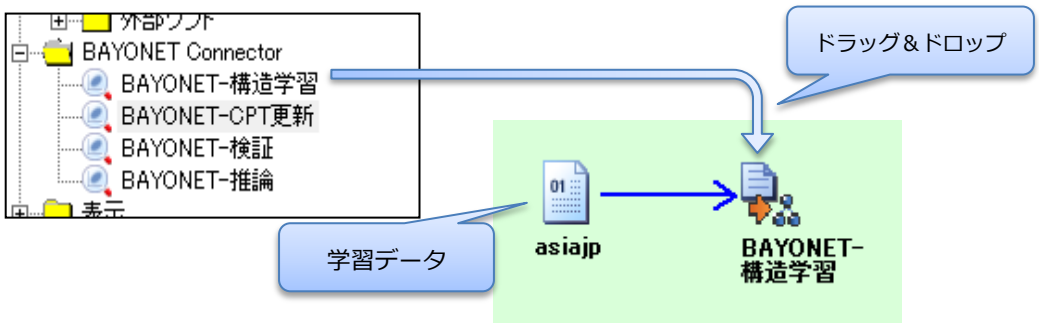
図 4　推論ツール（Excel アドイン）

4. BAYONET と Visual Mining Studio の連係

BAYONET バージョン 6.1（2013 年 9 月リリース）より Visual Mining Studio と連係ができるようになりました。Visual Mining Studio はビジュアルプログラミング環境を提供しており、データの加工や最新の分析手法を多数アイコンとして実装しています。このアイコンをドラッグ＆ドロップし、線で結ぶことによって分析フローを定義します。

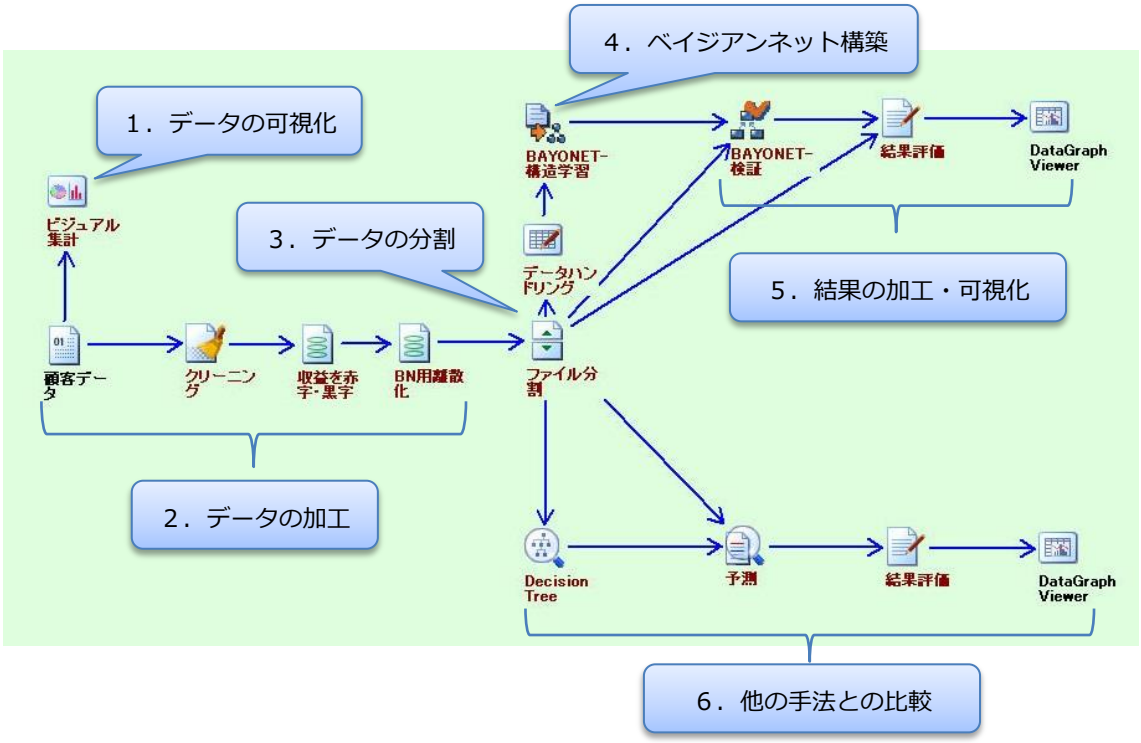
Visual Mining Studio がインストールされている PC に BAYONET をインストールすると、BAYONET の構造学習、CPT 更新、推論、検証の四つの機能が Visual Mining Studio に追加されます。CPT とは条件付き確率表の略で、ベイジアンネットワークのパラメータです。

- 構造学習
  - ✓ 構造学習アイコンをクリックすると BAYONET が起動しますので、構造学習ウィザードを使ってモデル構築を行います。
  - ✓ 学習データから構造学習アイコンへと矢印を引くことで学習データを指定します。

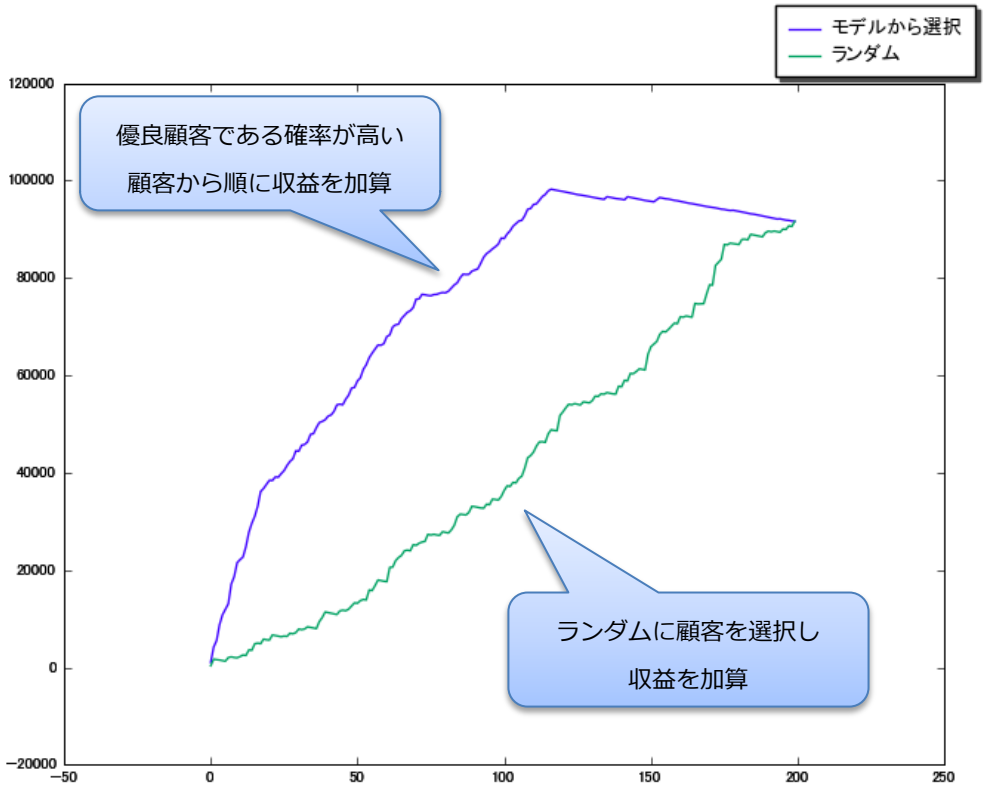


- CPT 更新
  - ✓ 構造学習アイコンで構築したモデルを、別の学習データで CPT を更新します。
  - ✓ 完全データだけでなく不完全データを使って CPT を更新することができます。不完全データの場合は EM アルゴリズムを用いてパラメータを推定します。
- 推論
  - ✓ 構造学習アイコンで構築したモデルを使って、入力データの欠損を推定値で補完します。複数のカラムを同時に補完することができます。
- 検証
  - ✓ 構造学習アイコンで構築したモデルの検証を行います。
  - ✓ 目的変数の適合率、再現率などを算出します

次の分析フローは構造学習と検証の二つの機能を使って、顧客データから、優良顧客かどうかを判定するベイジアンネットモデルを構築し、その性能を可視化しています。



1. データの可視化ではビジュアル集計機能を使い、各カラムの度数分布、クロス集計を見ながらデータの性質を把握しています。
2. データの加工では欠損を含むデータの除去、ベイジアンネット構築のために連続変数を離散化してカテゴリ変数にしています。
3. データの分割では、データを学習データと検証データに分割しています。
4. ベイジアンネット構築では、加工されたデータを使ってベイジアンネットモデルを構築しています。構造学習アイコンをクリックすると BAYONET の GUI が起動しますのでスタンドアロン版と同様にモデル構築を行います。
5. 結果の加工・可視化では検証結果を使って、収益を計算・描画し、モデルの性能を可視化しています。



6. 他の手法との比較では決定木による優良顧客判定モデルを構築し、ベイジアンネットとの性能を比較しています。

このように Visual Mining Studio との連係によって、ベイジアンネットによる分析の可能性が大きく広がります。