大量データから知識を抽出するベイジアンネットワークの学習技術とその応用

営業本部 事業開発部
鈴木 恒一

1. はじめに

現在、様々な情報が電子化され、企業などには大量のデータやログなどが蓄積されるようになった。またRFID やセンサーの普及に伴ってリアルタイムに大量のデータが生成されている。そのため、これらの大量データを効率よく分析して知識を抽出し、業務に生かすための技術が求められている。

事業開発部に属する当社新技術開発チームでは、大量データから知識を抽出する機械学習技術に注目し、ベイジアンネットワークの研究を電気通信大学と共同で行っている。大量データを効率よく処理する独自の手法を考案して学習エンジンを開発し、複数の業務への応用を展開し始めている。

本稿では、2章でベイジアンネットワークと学習技術の紹介をし、3章で我々のチームの保有技術と提供可能機能の説明を行う。4章では、各部門と協力して展開している業務への適用例を紹介する。

なお「機械学習」は、計算機を用いてデータから統計的性質を抽出し、確率分布の推定や予測のための統計モデルを自動的に構築する技術の総称である。なかでも本稿では、ベイジアンネットワークモデルの学習について述べる。

2. ベイジアンネットワークと学習技術

ベイジアンネットワークは、物事を確率的に推論する「確率推論」技術の一つである。我々の保有技術は、大量データからベイジアンネットワークを効率よく構築する学習技術であるが、まず確率推論とベイジアンネットワークについて説明を行う。

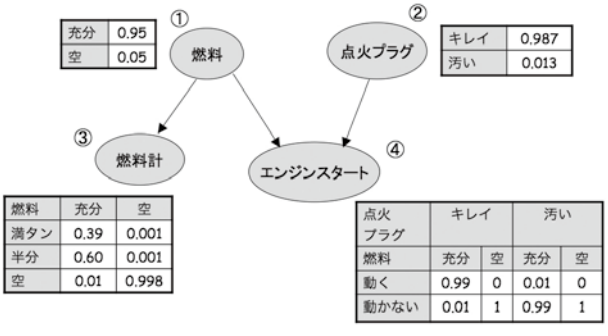
(1) 確率推論

確率推論は、「Aという事象が起きた時にBという事象が起こる確率は○%である」のように確率的に推論を行う技術である。「AならばBである」と決定論的に導かれる論理推論とは違い、不確定要素の多い事象のモデル化に向いている。「熱がある。喉が痛い」などの症状からその病因を探る原因推定や、「売上、利益、資産、借り入れ状況」から倒産の確率を推論する予測など、様々な分野に適用可能である。

(2) ベイジアンネットワーク

ベイジアンネットワークは、それぞれの事象をノードで、その関連を向きを持つリンクである有向リンクとして表現した非循環有向グラフ(ループを含まない有向グラフ)で、それぞれの事象の起こる確率をノードごとの確率テーブルとして持っているものである。図1にエンジンの故障診断を行う非常に簡略化されたベイジアンネットワークの例を示す(※1)。4つの円が事象を示すノードであり、数字の入った表がそれぞれのノードが持つ確率テーブルである。右下のノード④は「エンジンがかかるかどうか」を示し、「燃料があるかどうか」を示すノード①と「プラグが汚れているかどうか」を示すノード②の影響を受ける。④の確率テーブルには、燃料がある時とない時、プラグがきれいな時と汚れている時にエンジンがかかる確率がそれぞれ収められている。ベイズの定理から、エンジンがかからない時に燃料が空である確率とプラグが汚れている確率を計算することができ、エンジンの故障の原因推定をすることができる。

図1 エンジン故障診断の例



(※1) Jensen, F. V. and Nielsen, T.D.: Bayesian Networks and Decision Graphs, Springer Verlag, 2nd edition 2007.

ベイジアンネットワークは、(a)様々な事象をノードとして表現できること、(b)モデルの理解が容易なこと、(c)全てのノードを入力にも出力にも使えること、(d)他の推論技術に比べて推論精度が高いことから、非常に柔軟で使いやすく適用範囲の広い技術であり、応用例も数多く報告されている。

(3) ベイジアンネットワークの学習

では次に、ベイジアンネットワークの作成手法に目を向けてみる。図1で示したベイジアンネットワークでは因果関係が明確であり、モデルを作成するのは非常に簡単である。また確率テーブルも、過去の故障報告などからエンジンがかからなかった場合の原因として燃料がなかった場合、およびプラグが汚れていた場合の報告事例を集計することで容易に確率を求めることができる。しかしながらこのような単純なベイジアンネットワークでは実用とすることはできないであろう。

さらに多くのノードを持つ複雑なベイジアンネットワークを作成する場合、それぞれのノード間の関連性を判断することは難しくなる。また接続されるノードが増えるに従って次元が増える確率テーブルを全て埋めることは困難な作業となり、推論精度の高いベイジアンネットワークを人手で作成することはすぐに非現実的となる。

そこで注目されるのがデータからノード間の関連性を計算して適切に接続し、データを数え上げて確率テーブルを埋める学習技術である。ところが従来の学習技術では、ノード数やそれぞれのノードが取りうる状態の数が増えるに従って指数関数的に計算時間が増えてしまうという課題を抱えている。通常のデスクトップレベルの性能のPCでは、数百ノード程度までの学習が限度となる。

3. 保有技術と提供可能機能

当社新技術開発チームでは、森下研究員を社会人大学院生として電気通信大学へ派遣し、大規模ベイジアンネットワークの高速学習をテーマとして植野真臣准教授と共同研究を行ってきた。大規模なベイジアンネットワークを高速に学習するアルゴリズムTPDA (Three Phase Dependency Analysis)(※2) を改良してさらに高速化した手法(※3) を考案し、独自の学習エンジンを実装した。図2にこのエンジンの学習結果を示す。37ノードのベイジアンネットワークの学習において、サンプルデータ数を増やした時の計算時間と計算時間のばらつきを示す標準偏差(SD)のグラフである。提案手法(Proposed)では既存手法のTPDAよりも高速に、かつ計算時間のばらつきを非常に抑えて学習することができている。計算時間のばらつきが少ないということは、学習時間の見積りがしやすいということである。また図3のノード数を増やした時の計算時間のグラフに示すように、従来手法では500ノード程度までしか学習できない計算環境で、1000ノードのベイジアンネットワークの学習を実用的な時間で行うことができている。

図2 計算時間と標準偏差

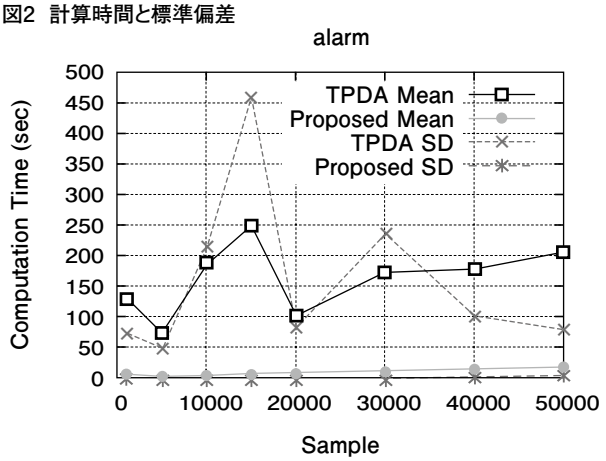
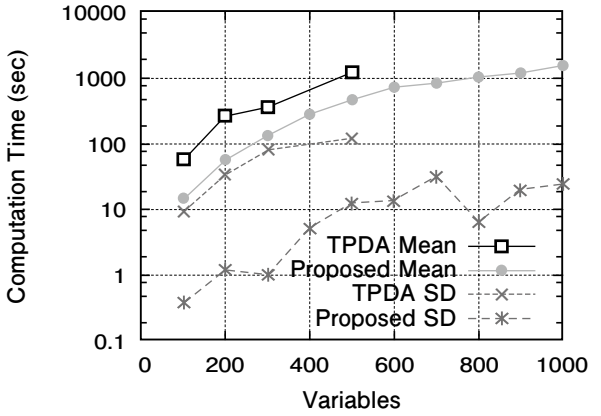


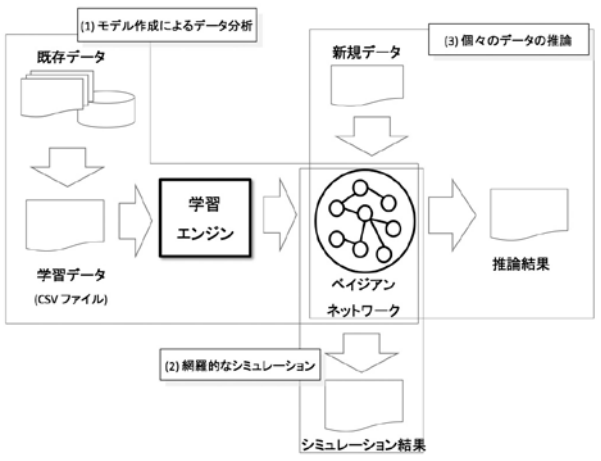
図3 ノード数と計算時間



次章で紹介する実際の適用例ではまだ数十ノードレベルのベイジアンネットワークしか計算していないが、既存手法よりも高速でかつ学習時間の予測が容易であるという強みは非常に有効である。

この学習エンジンを核として、我々のチームでは以下のような機能を提供することが可能となっている(図4)。

図4 提供可能機能



(※2) Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W.: Learning Bayesian networks From data: an information-theory based approach, Artificial Intelligence, Vol.137, No. 1-2, pp. 43-90 (2002)
(※3) 森下民平, 植野真臣: ベイジアンネットワーク学習アルゴリズムTPDAの高速化,人工知能学会全国大会(第24回)論文集, 2010

(1) データを学習してグラフモデルを作成し視覚化する

例えば顧客データや購買履歴、アクセスログなど既存データから学習データ(CSVファイル)を作成し、学習エンジンでグラフモデルを構築することにより、どのノード(年齢、性別、閲覧ページ、購買の成否など)間の関連が強いかを読み取ることができる。購買の成否と関連のあるノードが明確になれば、それに対する対策を立てることができるであろう。ただしノード数が増えるに従ってベイジアンネットワークは複雑化し、知見を読み取るのが難しくなるため、(2)のシミュレーションによって有用な情報を抽出する必要がある。またモデル作成の目的を明確化して生データから必要な学習データを作成する前処理には業務知識が必要となるため、業務を深く理解するエキスパートとの連携が必要である。

(2) グラフモデルをもとに網羅的なシミュレーションを行う

(1)で作成したグラフモデルをもとに、知りたい情報を持つノード(購買の成否)と関連のあるノードが持っている状態を網羅的に組み合わせて確率推論エンジンで処理することにより、どのような条件の下で購買確率が高くなる(もしくは低くなる)のかをシミュレーションすることができる。購買確率の高い条件を備えているが未購入の顧客であれば、購入を期待できるであろう。

(3) 新規データをグラフモデルに入力し、個々のデータについての推論を行う

(1)で作成したグラフモデルをもとに、学習では使用しなかった顧客リストを確率推論エンジンで処理することにより、どの顧客の購買確率がどのくらいあるかを推論することができる。これにより購買確率の高い顧客からコンタクトを行うことで、効率的に営業を行うことができるであろう。

4. 実業務への適用例

現在、大きく分けて二つの領域への適用を進めており、関連する部門と連携して活動を行っている。一つはCRM領域への適用で、顧客データ分析による売上/利益向上への貢献である。既に当社顧客と2011年の実用化を前提とした実証を開始しており、主にWebサイトのアクセスログデータを受領して分析を行っている。いくつか結果は出つつあるが、実証中なので本稿での報告は差し控えることとする。また金融分野において既存顧客深耕のための商品推薦や個人ローン向けの与信モデル作成などのテーマについて適用の検討を始めている。もう一つは医薬安全性領域への適用で、薬と副作用の相関関係分析である。米国FDA(食品医薬品局)が公開しているデータを取得し分析を行い、それをもとに複数の製薬会社と情報交換を実施している。以下では医薬の事例を紹介する。

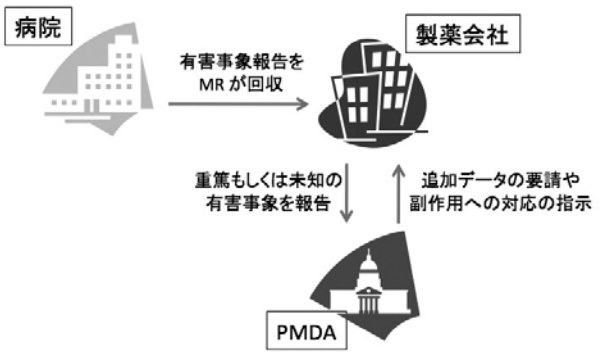
4.1 市販後薬剤の有害事象分析(シグナル検出)

現在、当社医薬BTO部門と連携して活動しているのが薬と有害事象の関連性分析である。ある薬と複数の併用薬、そしてそれらの薬を服用した時の有害事象の報告データから、薬と有害事象の関連性や、複数の薬を飲み合わせた時の有害事象の起こりやすさなどを計算することができる。ここで「有害事象」とは、薬との因果関係のあるなしに係わらず、薬が投与されたのちに患者に発生した様々な負の事象である。この中からある薬との因果関係が認められたものを「副作用」と呼ぶ。副作用が疑われるような有害事象をデータから検出することを「シグナル検出」と呼ぶ。

(1) 製薬会社における分析ニーズ

新薬が承認され市販された後に起きた有害事象の報告は、病院から MR を通して各製薬会社に集められる。各製薬会社では収集された報告を分析し、死亡などの重篤な事象や未知の有害事象をPMDA(医薬品医療機器総合機構)へと報告を行う。PMDAでは、各製薬会社から上がってくる報告をさらに分析し、注目すべきデータがあった場合には製薬会社に追加のデータを要求し、薬と副作用の関連性を見極め、薬の添付文書改訂などの対応を指示する(図5)。製薬会社では、PMDAからの予期せぬ問い合わせへの対応が非常に負担となってきているため、問い合わせなどにいつでも対応できるよう、薬との因果関係が疑われるような事例をあらゆる角度から分析を行って把握しておきたいニーズがあり、我々の分析手法もその支援の一つとなる。

図5 有害事象の報告経路



(2) 既存手法ではできない多剤組み合わせにおける定量的分析

シグナル検出に使用されるデータマイニング手法には、これまでにもいくつか存在する。各国の規制当局がそれぞれ異なる手法を採用しており、日本でも2009年度からデータマイニング手法によるシグナル検出の運用が始まっている。しかしながら既存のデータマイニング手法は、ある薬と有害事象のペアに注目し、その報告事例数をその他の事例数と比較

して有意に多いかどうかを検出するもので、複数薬の組み合わせの相互作用による有害事象を分析することはできない。それに対してベイジアンネットワークに基づく手法を用いることで、複数薬の組み合わせによる有害事象の発生確率を効率的に計算することができる。有害事象の起きやすさ/起きにくさを、確率値として定量的に計算することができるのも強みの一つである。自社の薬が他社のライバル薬と比べて有害事象が起こりにくいことを定量的に示すことができれば、販売促進にも大きな力となる。

(3) 技術検証

ベイジアンネットワークに基づく手法を技術検証するために、米国FDAにて公開されている過去5年分の報告データ(AERS)223万件を利用して、ベイジアンネットワークの学習を行った。AERS中で報告されている薬剤は約13,000種類、有害事象は約6,300種類ある。ここから前処理として、ある注目薬Aの含まれる報告の中から併用薬と有害事象を抽出し、45薬剤の学習データを作成した。この学習データは、ある有害事象に関して服用された複数の薬がそれぞれ「飲んだ(T)/飲まない(F)」の2値で表されたCSVファイルとなっている。この学習データを我々の学習エンジンで処理することにより、ある有害事象Xと関連の強い11薬剤がリンクで接続されたモデルが作成でき、関連図を描くことができた(図6)。ここで学習したモデルを推論エンジンにかけることで、それぞれの薬が飲まれた時に有害事象が起こる確率をシミュレートすることができる。全ての組み合わせについて網羅的に推論することによって、何も条件が与えられてない状況に比べて有意に有害事象が発生する組み合わせを抽出することができる(表1)。なお発生確率とは、この有害事象Xの報告中にこの薬剤の組み合わせが現れる確率であり、この薬剤を飲んで有害事象X が発現する確率ではない。

図6 有害事象Xと薬剤の関連図

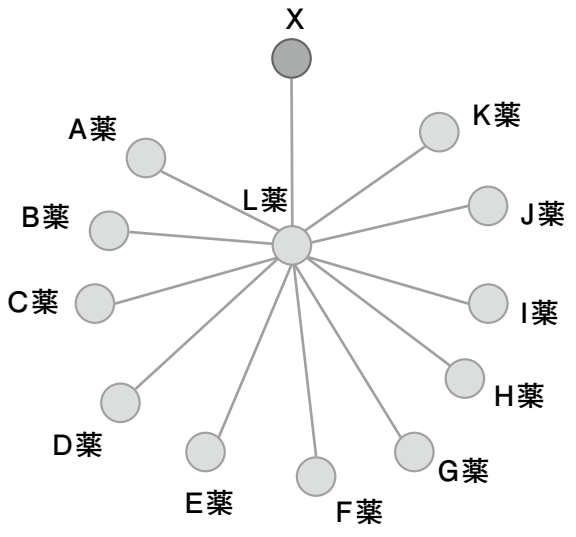


表1 薬剤の組み合わせと発生確率の推論値

薬剤の組み合わせ	発生確率
G, H, I, J, K	33.33%
A, B, C, D	25.00%
C, G, K, H	25.00%
C, G, E	23.08%
F, G, L	14.89%
:	:
条件なし	0.60%

(4) 独自エンジンの強み

このAERSのデータは、全体としては膨大な報告の中で注目薬の報告が現れる割合はごくわずかであり、極端に偏ったデータとなっている。通常のベイジアンネットワークの学習エンジンでは、このようなデータの学習は非常に困難である。我々はこのような問題向けに学習エンジンを調整し、極端に偏ったデータでも妥当なベイジアンネットワークの作成を可能とした。これは我々が理論からエンジンの実装までの知識/技術をカバーしていることから実現できていることである。

(5) 今後の動き

現在すでにAERSデータの分析結果をもとに、いくつかの製薬会社にベイジアンネットワークを利用したシグナル検出手法の紹介と社内データ分析トライアルの要望ヒヤリングを行っている。今後は日本の有害事象報告データを分析して、引き続き製薬会社にアプローチしていくとともに、論文発表などを通じてアカデミックな実績を積み、PMDAや製薬会社に対して多剤での分析の必要性を訴求していく予定である。

5. おわりに

ここまで、ベイジアンネットワークの学習技術の説明から適用事例紹介まで行ってきた。この技術を応用したビジネスのイメージは掴んでいただけたのではないかと思う。技術的に興味深い課題はまだいくつもあるものの、ビジネス化の兆しが見えてきていると考えている。このベイジアンネットワークの学習技術は理論部分からエンジン実装部、業務ごとのアプリケーション部分まで全て当社がカバーしているため、ビジネス展開時にも非常に強い競争力を持つと考えている。この競争力を維持するための基礎研究部分を地道に継続しつつ、ビジネス展開を続けていく考えである。