

一般口演

一般口演7

医療データ分析3（レセプトデータ・治験）

2018年11月23日(金) 14:20 ～ 15:50 H会場 (福岡サンパレスHパレスホール)

[2-H-2-3] 臨床レジストリを活用するための患者群状態モデルの構築

○石井 雅通¹, 美代 賢吾¹, 杉山 雄大², 大杉 満², 植木 浩二郎² (1.NCGM (国立国際医療研究センター) 情報基盤センター, 2.NCGM (国立国際医療研究センター) 研究所 糖尿病研究センター)

【はじめに】当センターでは日本糖尿病学会との共同事業である「診療録直結型全国糖尿病データベース事業（Japan Diabetes comprehensive database project based on an Advanced electronic Medical record System:J-DREAMS）を推進しており、10～20万人規模の登録をめざしている。今後、収集した臨床データを活用した人工知能システムの構築などが期待されている。【課題】登録された臨床データは300項目以上であり、数万規模の多施設患者の状態を直観的に把握することは困難である。データ活用にあたっては従前よりクリニカルクエスチョンごとの個別統計処理が不可避であった。また臨床データを活用して効果的な人工知能システムを構築するには臨床専門家のドメイン知識を効率的に反映可能なデータプレパレーション手法が求められている。【提案手法】蓄積された多施設臨床データを効果よく汎用的に活用する手法として人工知能技術のひとつであるベイジアンネットワークにより構築した患者群状態モデルの利用を提案する。【検証方法】専門医が入力したレジストリ収集対象項目を特徴量として、糖尿病患者に特徴的な検査結果値の変化を正解ラベルとした患者群状態モデルのプロトタイプを構築し、専門医による評価を実施した。【結果及び考察】ベイジアンネットワークによる患者群状態モデルを構築することで患者群の状態の可視化が実現した。モデルを探索的に利用することで専門家による気づきの機会や検索条件の組み合わせによる暗黙知のフィードバックの機会を提供することが可能となり、持続的に知識ベースを発展させるシステムの実現可能性を示すことができた。

臨床レジストリを活用するための患者群状態モデルの構築

石井 雅通^{*1}、美代 賢吾^{*1}、杉山 雄大^{*2}、大杉 満^{*2}、植木 浩二郎^{*2}

*1 NCGM(国立国際医療研究センター)情報基盤センター、*2 NCGM 研究所 糖尿病研究センター

Development of patient state model to overview clinical registry database

Masamichi Ishii ^{*1}, Kengo Miyo ^{*1}, Takehiro Sugiyama ^{*2}, Mitsuru Ohsugi ^{*2}, Kohjiro Ueki ^{*2}

*1 NCGM (National Center for Global Health and Medicine) Healthcare Intelligence Technology Center,

*2 NCGM Research Institute Diabetes Research Center

[Introduction] NCGM is promoting the Japan Diabetes comprehensive database project based on an advanced electronic medical record system (J-DREAMS) which is a joint project with the Japan Diabetes Society, aiming to register between 100,000 and 200,000 people. In the future, development of an artificial intelligence system using clinical data collected by J-DREAMS is expected. [Problem] Registered clinical data is more than 300 items and it is difficult to intuitively grasp the state of multi-institutional patients of tens of thousands. For data utilization, individual statistical processing for each clinical question has been unavoidable. [Proposed method] We propose a patient-state model expressed in the Bayesian network which is one of artificial intelligence techniques as a method to effectively utilize accumulated multi-institutional clinical data in general purpose. [Result] A prototype of the patient-state model was developed, and evaluations were carried out by a specialist. [Discussion] By building a patient-state model by Bayesian network, patient condition was visualized. By facilitating database search, we made it easy for experts to provide his awareness and to feedback his tacit knowledge. [Conclusion] We showed that clinical registry database can be visualized by constructing patient-state model with Bayesian network architecture, and it can be helpful for constructing clinical knowledge base.

Keywords: Artificial Intelligence, Bayesian Network, Clinical Question, Knowledge base

1. はじめに

当センターでは日本糖尿病学会との共同事業である「診療録直結型全国糖尿病データベース事業 (Japan Diabetes compREhensive database project based on an Advanced electronic Medical record System : J-DREAMS)」を推進しており、10~20 万人規模の登録をめざしている。今後、収集した多施設の臨床データを活用することにより、観察する臨床データ項目の標準化による診療の質の改善、合併介入の同定、糖尿病薬の副作用についての発生頻度及びリスク因子の確認、症例の未解決課題の発見といった成果を見込んでおり、さらには臨床データから抽出した臨床知識を搭載した糖尿病領域における人工知能システムへ発展させることが期待されている。¹⁾

J-DREAMS を始めとする多施設参加型の患者レジストリ事業においては当該疾患領域を専門とする医学会が標準となる観察項目を予め定めたうえで当該疾患領域の臨床専門医が登録した観察項目データを各施設から収集して構築する方式が一般的に実施されている。この方式にて収集されるデータ項目は専門家による判断のフィルターを通した結果であるから、人工知能技術活用の視座からは専門家による正解ラベルとなりうる。また学会選定の観察項目それ自体が当該領域の臨床経験の蓄積から導き出されたドメイン知識であり、主要な特徴量となりうる項目で構成されている。

J-DREAMS 事業では糖尿病専門医による観察項目と同時に、各施設の電子カルテシステムに蓄積された対象症例に関する患者基本情報、投薬情報、検査結果情報といった様々な臨床情報を厚生労働省標準である SS-MIX2 標準化ストレージを経由して時系列の構造化データとして収集しており、その活用が期待されているところである。²⁾ しかしながら、人工知能技術等を適用したデータ駆動型のデータ活用に

あたっては、その準備として臨床データ特有の欠損値へ対応に始まり、新 NP 問題、スパースネス問題に直面することになる。³⁾⁴⁾

1.1 臨床データの欠損値対応

一定の成果を出している人工知能技術として深層学習アルゴリズムが挙げられるが、その精度は学習データの質・量に依存する部分が多い。センサー機器等のIoT(Internet of Things)から取得される大量の連続的な時系列データと相性がよい。一方、電子カルテシステムから収集可能な臨床情報は様々な要因によりデータ発生が一定ではないことがあり、結果的に欠損値が多く発生する。

欠損値への対応にあたっては時間軸の評価尺度を変更するなどして離散化する、当該のデータ項目の特性に応じたアルゴリズムにより新たな統計値(平均値、ばらつき度、単位時間あたりの変化量など)を特徴量として算出するなど、データ特性に応じた工夫が求められる。臨床データから適切な特徴量を導出して学習用データを準備できるか否かで最終的な学習モデルの品質が左右されることになる。

1.2 臨床データの新NP問題

電子カルテシステムを導入している多くの総合病院規模の医療施設において、採用医薬品は 2,000~3,000 品目、検査結果値は 3,000~5,000 項目、観察項目はその数倍あり、全体のデータ項目数としては 10,000 を超える。さらに今後ゲノム医療の進展により患者の遺伝子変異を取り扱うことが見込まれており、その場合のデータ項目数は数万規模となる。

一方、患者の疾患は多種多様であり、特定の疾患に対してデータ駆動型解析を行う場合は対象症例件数が絞り込まれることになる。特に対象が稀少疾患となると一施設あたり数件程度のデータ件数となることもある。一般に母数が多い疾

患であると考えられる「がん」を扱う場合も、精度のよい学習モデルを作成するにはがん種や進行度などの条件ごとに学習データの準備が必要となりデータ件数が絞り込まれる。データ件数(n)に対応するパラメータ数(p)が膨大である($n < p$)という新 NP 問題に常に悩まされる。

1.3 臨床データのスパースネス問題

新 NP 問題に加えて、臨床データのスパースネス問題がある。説明変数候補となる臨床データは多種多様であることに加えて、説明変数ごとの実測データは欠損値が多く存在し、全体を俯瞰するとそれぞれの臨床データはまばらに存在するという特徴(スパースネス(スパース性))をもつ。そのためデータ駆動型アプローチでデータ活用するには、スパースネス問題(ゼロ頻度問題ともいう)への対策が必要となる。主たるアプローチとしては主成分分析による次元削減が挙げられる。

2. 目的

臨床レジストリデータベースから臨床知識を抽出して人工知能システム上で活用することを目指しているが、実現にあたっては臨床知識の源泉となる“学習データ”を人工知能学習に適した状態で準備する必要がある。学習データの準備とは収集されたデータ項目から学習に適切な特徴量データを導出することであるが、その際には臨床のドメイン知識に加えて臨床レジストリデータベースに収集された症例データの特徴理解が前提となる。

そこで本論では臨床レジストリデータベースの症例データの特徴理解をサポートする技術開発を目的として、臨床レジストリに登録された症例群を可視化して特定の患者群の状態を表現する“患者群状態モデル”の構築を試みる。

2.1 従来のアプローチ

臨床現場では患者本人に相対する、電子カルテに記録された臨床情報を参照するなど患者状態を把握することになるが、そうした症例が数万件蓄積された臨床レジストリデータベースに対して全ての症例データを参照するアプローチを取ると相応の時間コストが必要となる。

そこでデータベースに登録された症例の傾向や特徴を把握するために記述統計を活用することになるが、その場合は臨床クエスチョンの数だけ分析条件が異なるため都度データを抽出し統計算出するアプローチとなる。この方法は既にデータの特徴が把握されて分析視点が定まっている場合は効率が良いが、分析視点自体が探索対象である場合は統計算出を繰り返すため効率が悪い。

分析視点を変更しながら探索するアドホックなデータ分析手法は BI ツールとして提供されているが、例えば J-DREAMS データベースを BI ツールで効果的に分析するには 50 以上のテーブル構造及び 300 以上のデータ項目の関連を事前に理解しておくことで効果的な活用が可能となるが、臨床専門家が実践するには負担が大きい。

2.2 グラフ表現モデルによる可視化

従来の手法は、分析軸ごとのデータ集計を繰り返す手法でありデメリットがある。そこでベイジアンネットワークを活用した可視化により簡易化を試みる。

ベイジアンネットワークは分析軸の要素となるデータ項目とデータ構造をグラフ表現するモデルであり、直感的に症例データを取扱うことができる。具体的には、J-DREAMS のデータ項目に適用することで 300 以上のデータ項目とデータ項目間の関連がグラフ構造で可視化されて全体を俯瞰しつつ、任意

のノード(データ項目)に事前確率として特定の値を指定することで他のノードの事後確率、すなわちデータ件数を求めることができる。この機能により、症例データの集合である臨床レジストリデータベース全体から特定の条件を満たす症例の集合体(ここでは、患者群と呼ぶ)を抽出して表現するための“患者群状態モデル”が提供できると考えた。

3. 方法

当センターから J-DREAMS に提出した症例データを臨床レジストリデータベースの実例として、患者群状態モデルのプロトタイプを以下の方法で構築する。

- (1) データ項目の抽出
電子カルテシステムのテンプレート機能に実装している「J-DREAMS 糖尿病標準診療テンプレート」(以下、テンプレートとよぶ)で入力対象となっている観察項目を全て抽出する
- (2) ベイジアンネットワークモデル定義
(1)で抽出した観察項目をノードとして定義する
テンプレートにおける入力階層に従い、ノードの階層配置を行う
ノード間の矢印は上位階層の項目を親ノード、下位階層の項目を子ノードとして定義する
- (3) 症例データ抽出
当センターから J-DREAMS に提出後に返却されるリレーショナル・データベースから(1)で抽出したデータ項目に対応する症例データを抽出する
- (4) 学習データ作成
(3)の症例データを入力として学習データを作成する
 - i データ項目を検査し、クレンジングする
 - ii 時系列を考慮した特徴量データを作成する
例) 前回検査値からのカテゴリ変化、変動量等
 - iii 数値データをカテゴリ変数に変換する(離散化)
例) 実年齢を年代カテゴリに変換
- (5) ベイジアンネットワークの作成(パラメタ学習の実行)
(4)で作成した学習データから(2)で定義したベイジアンネットワークモデルの各ノードの条件付き確率分布を求める

4. 結果

作成したベイジアンネットワークのモデルを図1に示す。

患者群状態モデルのプロトタイプを構築し、糖尿病専門医による評価を実施した。主な評価を以下に示す。

- 従前の記述統計と比べてテンプレート構造が可視化されており全体が俯瞰できて良い
- 任意入力項目に対する入力の偏り状況が把握できて良い
- 事前確率条件を変更することで探索的なデータ検索が可能となった
- 診察中患者状態を事前確率に設定し事後確率を示すことができれば臨床現場における活用の可能性がある

5. 考察

臨床レジストリデータベースにベイジアンネットワークを活用して患者状態モデルを構築することで、患者群のデータ状態を可視化して全体を網羅的に俯瞰することが可能となり、探索的に患者群の傾向を分析する手段を提供することができた。

ベイジアンモデルの利用により当該症例の専門家自身が臨床クエスチョンに基づき臨床レジストリデータをスクリーニングに利用することで、専門家による気づきの機会や検索条件の組み合わせによる暗黙知のフィードバックの機会を

提供できる可能性がある。フィードバックをデータベース収集項目へ反映することによりデータ活用の好循環が期待できる。

今後は、時系列を考慮した特徴量データを増やしていくことで臨床専門家の臨床的クエリに十分応えられるモデルへと成長させていきたい。本取り組みにより J-DREAMS 事業が目指している臨床データベースを活用した i) 診療の質の改善 ii) リスク因子の同定 iii) 未解決課題の発見に貢献できると考えている。

6. 結論

臨床レジストリデータベースからベイジアンネットワークによる患者群状態モデルを構築することにより、臨床レジストリの可視化を実現し、データベース利活用による臨床知識ベース構築に向けた一助となりうることを示した。

謝辞: この研究の一部は AMED の課題番号 JP17lk1010010 の支援を受けた

参考文献

- 1) 国立研究開発法人国立国際医療研究センター, 一般社団法人日本糖尿病学会. 診療録直結型全国糖尿病データベース事業, 2017. [http://jdreams.jp/ (cited 2018-Aug-21)].
- 2) 小南亮太, 美代賢吾, 杉山雄大ら. 診療録直結型全国糖尿病情報データベース(J-DREAMS)の構築と多施設データ統合の諸課題. 第36回医療情報学連合大会論文集 2016 ; p292-293.
- 3) 統計関連学会連合理事会および有志. 統計関連学会連合理事会および有志, 我が国の統計科学振興への提言, 2007. [http://www.jfssa.jp/wp-content/uploads/2017/01/TokeiKagakuShinkou0702.pdf (cited 2018-Aug-21)]
- 4) 田中博. 人工知能(AI)とビッグデータの医療と創薬への応用, 2013. [http://ngbr.com/img/file54.pdf (cited 2018-Aug-21)]

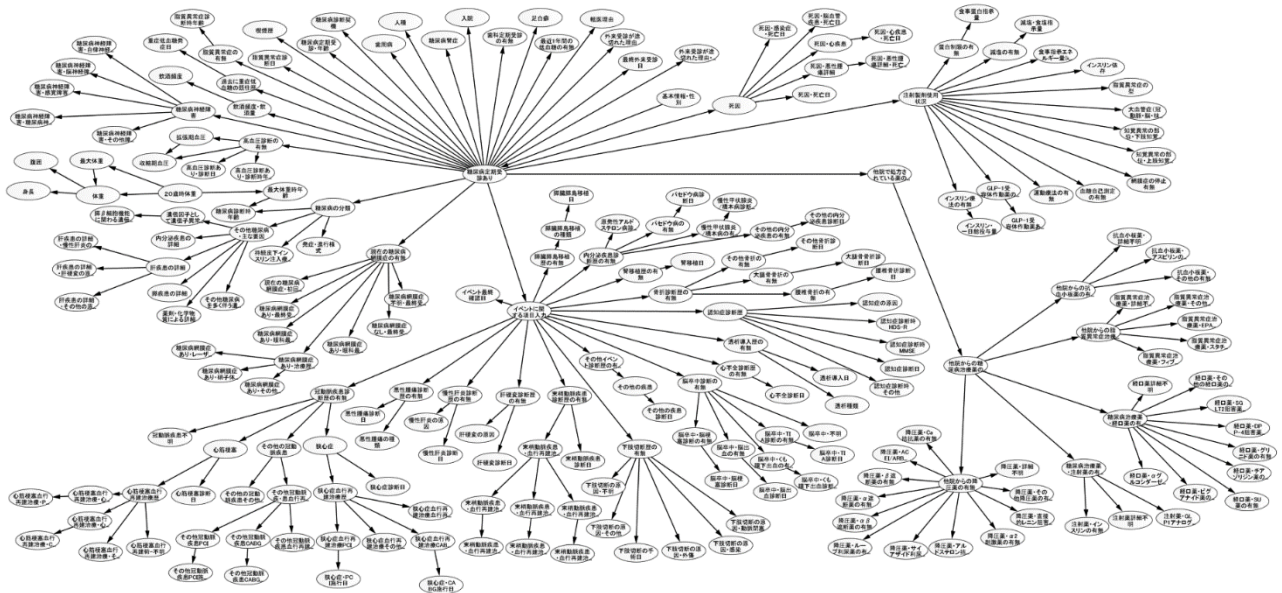


図1 ベイジアンネットワークのモデル(J-DREAMS データ項目より)

