これまでのおしごと

近原　鷹一

2019 年 6 月 20 日

1　教師あり学習に基づく時系列の因果推論

- **研究目的:**

  時系列の因果推論とは、入力として与えられた時系列データからその変数間の因果関係の有無・方向を出力する問題であり、知識発見を目的とするタスクである。時系列の因果推論は経済学、脳科学、バイオインフォマティクスなど、幅広い応用が期待される問題である。しかし、複雑な様相を呈する実データから高精度に因果関係の有無・方向を推定することは技術的に容易ではない。本研究課題では、高精度な時系列の因果推論技術を構築し、推定精度が要求される多くの実問題に応用することを目指す。

- **技術的背景:**

  時間依存する変数間の因果関係の定義はいくらかあるが、本研究課題では幅広く用いられている因果関係の定義である、Granger causality に着目する。これは、変数 X の過去の値が変数 Y の未来の値を予測するのに**有用**であれば、X は Y の原因である、として定義される。

  既存手法では、Granger causality を推定するために、予測の**有用性**を評価するための予測式として、(自己) 回帰モデルを用いる。回帰モデルを用いて、過去の X の値を用いた場合と用いなかった場合とで、未来の Y の値に関する予測誤差を比較し、過去の X の値を用いた場合のほうが予測誤差が有意に小さかった場合、X は Y の原因である，と推定する。

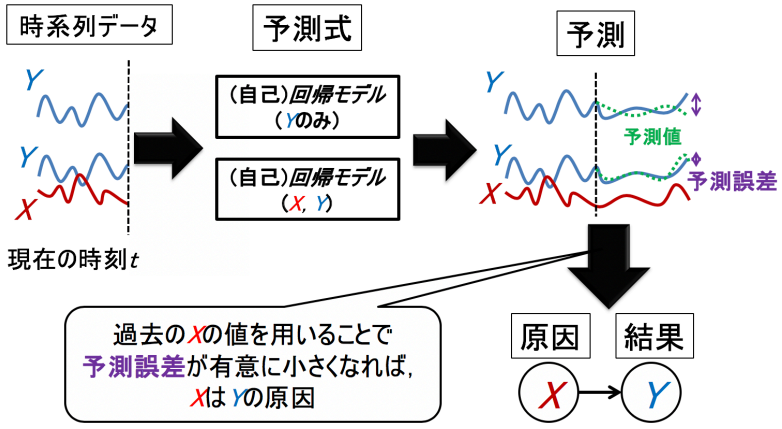

図1　既存手法による Granger causality 推定の概要

このような既存手法では、VAR モデルや一般化加法モデルといった回帰モデルを予測式として選択する必要があるが、選択した回帰モデルが所与の時系列データに対してうまく適合しない場合、Granger causality を正しく推定できないという問題点がある。データに対して適切にモデルを選択することは、データ分析に関する深い専門知識を要求されるため、これは実用上、重大な問題である。そこで本研究課題では、このようなモデル選択に関する問題を回避するべく、新たな Granger causality 推定のフレームワークとして、教師あり学習に基づく推定方法を提案する。

- **提案手法の概要:**
  提案手法では、入力として与えられる因果関係が未知の時系列データ（テストデータ）とは別に、因果関係が既知の時系列データ (訓練データ) を用意する。これは、因果関係が明らかな実データが入手可能な問題設定ではそれを用い、不可能な場合は人工データを用いる。提案手法では、このような訓練データを用いて分類器を学習し、学習した分類器を用いてテストデータの因果関係を推定する。提案手法は教師あり学習に基づいており、訓練データとして多種多様なデータが与えられている場合においては、高精度に Granger causality の有無・方向を推定することが期待できる。

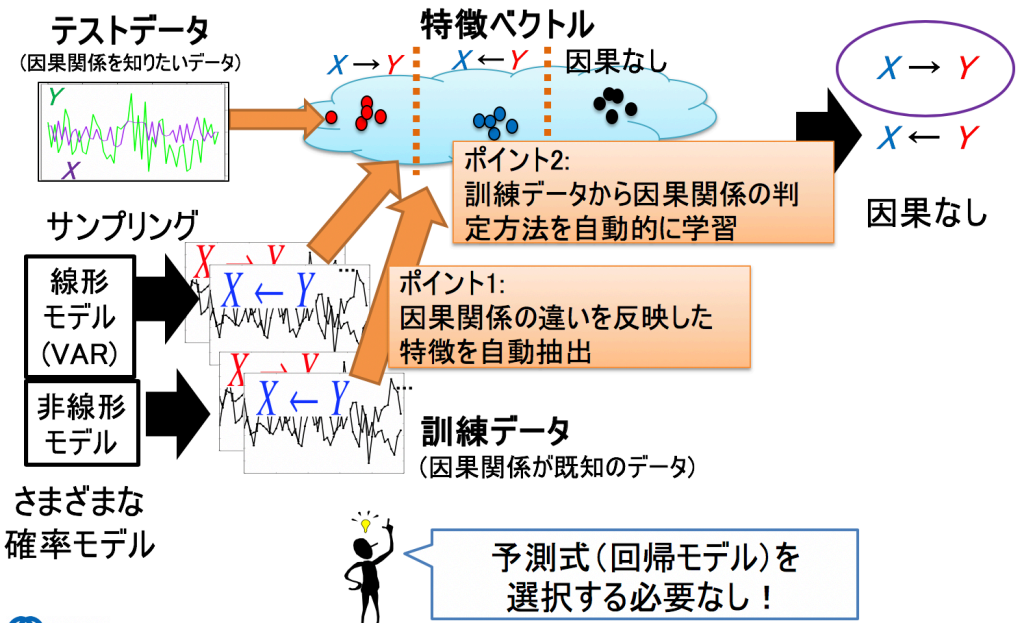

図2　提案手法による Granger causality 推定の概要

本研究では、Granger causality の有無・方向を表すクラスラベルを割り当てるような分類器を構築するため、Granger causality の定義に基づく特徴量表現を定式化した。一般に Granger causality は 2 つの条件付き分布が等しいか等しくないかで定義される。例えば変数 X が変数 Y の原因であるとは、過去の Y の値で条件づけられた未来の時刻の Y の条件付き分布と、過去の X, Y の値で条件づけられた未来の時刻の Y の条件付き分布とが等しくない場合として定義され、これらが等しい場合、X は Y の原因でないとされる。このような定義に基づき、提案手法では分布間距離尺度 MMD(maximum mean discrepancy) に基づいて条件付き分布間の距離を計算し、これを用いて特徴量を得る。

このような特徴量表現によって、Granger causality の有無・方向の違いによって十分異なる特徴量が得られることを実験的に確認した。また、提案手法が回帰モデルを用いる既存手法に比べて高い精度で Granger causality を推定することを、実験的に確認した。

- 主な研究成果:

  人工知能分野の最難関国際会議 IJCAI、情報処理学会論文誌 (TOM) に論文が採録され、国内研究会にて表彰、さらに招待講演も予定している。

  1. 査読付き国際会議:
     (a) Yoichi Chikahara, Akinori Fujino; "Causal Inference in Time Series via Supervised Learning", Proc. of the 27-th International Joint Conference on Artificial Intelligence Stockholm, Sweden, July 2018 (IJCAI2018; Acceptance Rate: 20%)
  2. 査読付き国内論文誌:
     (a) 近原鷹一, 藤野昭典, 教師あり学習に基づく Granger causality の推定; 情報処理学会論文誌数理モデル化と応用 (TOM), September 2018
  3. 表彰:
     (a) ベストプレゼンテーション賞; 情報処理学会数理モデル化と問題解決研究会 (MPS2018)
  4. 招待講演:
     (a) "Causal Inference in Time Series via Supervised Learning";
     情報処理学会 第１８回情報科学技術フォーラムトップコンファレンスセッション (FIT2019); September, 2019

2　無限関係モデルを用いたがん患者オミックスデータの統合解析

- **研究目的:**
  がん細胞では遺伝子に関する多種多様な異常が次々と蓄積されるため、がんという病気を理解し、治療法を確立することは困難を極める。これまでの研究から細胞のがん化はドライバー遺伝子と呼ばれる遺伝子によって生じることがわかっている。具体的には、ドライバー遺伝子の DNA コピー数 (複製数) に異常が生じ、その機能発現量の程度を表す mRNA 発現量に異常が発生し、これによりモジュールと呼ばれる下流遺伝子群の mRNA 発現量に異常を生じ、結果として次々と異常が蓄積されるとされている。蓄積された異常の程度によってがん細胞の形質は異なるが、類似した形質を示す細胞グループは一般にがんのサブタイプと呼ばれ、その特徴を理解することはがんを理解するうえで必要不可欠である。
  ドライバー遺伝子を発見し、がんのサブタイプを同定するためには、mRNA 発現量だけでなく、DNA コピー数などの他のオミックスデータを用いた統合的な解析が必要となる。本研究課題では、遺伝子に関する異種データを統合的に解析することで、ドライバー遺伝子を特定し、がんのサブタイプを同定し、ドライバー遺伝子が作用するモジュール群を推定するための手法を確立することを目指す。

- **技術的背景:**
  これまでの既存研究では、ドライバー遺伝子を推定することに特化した手法、がんのサブタイプの同定に特化した手法は提案されたが、両者を同時に行う研究は提案されていなかった。ドライバー遺伝子推定に関する既存手法では、各遺伝子の DNA コピー数と mRNA 発現量に基づいて遺伝子のクラスタリングを行い、両者の相関が高い遺伝子のクラスターをドライバー遺伝子の候補とした。がんのサブタイプの推定に関する既存手法では、がん細胞サンプルに関してクラスタリングを行い、類似した mRNA 発現量のパターンを示す細胞サンプルのクラスタを同定した。
  本研究課題では、遺伝子、がん細胞サンプル、モジュールに関するクラスタリングを同時に行う手法を提案する。同時にではなく逐次的にクラスタリングをする場合、例えば遺伝子、がん細胞サンプル、モジュールの順にクラスタリングを行うと、最初の遺伝子に関するクラスタリング結果に強く依存した結果になるという問題が生じるが、同時クラスタリングであればそのような問題を回避することができると期待される。

- **提案手法の概要:**
  本研究では、2006 年に Kemp 等によって提案された無限関係モデル (IRM) を拡張し、遺伝子・がん細胞サンプル、モジュール方向の同時クラスタリングを行うための生成モデルを提案する。
  元論文の IRM は離散値データを対象とした生成モデルであったが、連続値のオミックスデータを対象とできるよう、本研究ではモデルの拡張を行った。提案した生成モデルはノンパラメトリックベイジアンモデルであり、(1) 適切なクラスター数をデータから自動的に推定でき、(2) 生物学的なドメイン知識に基づいて事前分布を適切に設計することが可能である。後者に関して、例えば遺伝子の位置情報 (染色体座位に関する情報) を用いて位置的に近い遺伝子が同一クラスタに属しやすいよう事前確率を定めることで、位置情報を活用した遺伝子のクラスタリングが可能である。

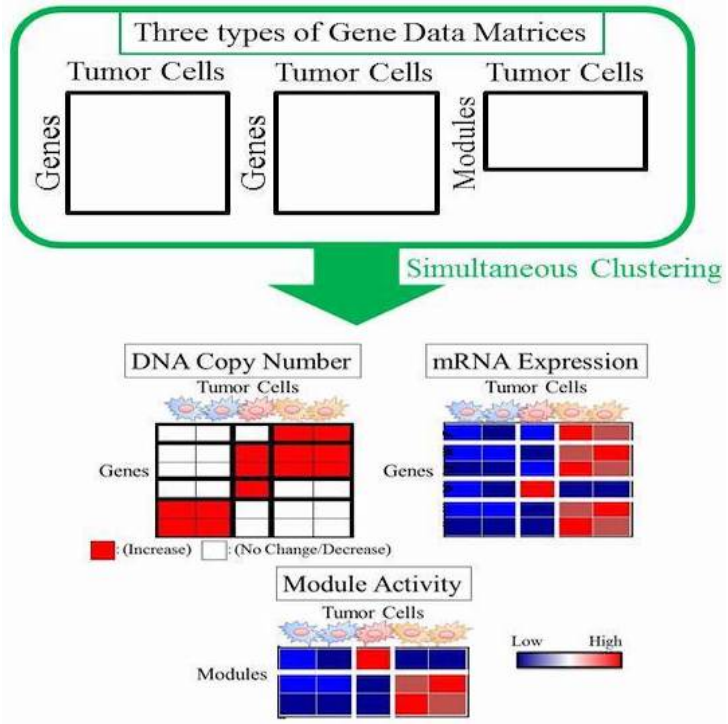

図3　提案手法の概要

このような特徴を持つ提案手法を、がんゲノムデータベース TCGA の大腸がん患者のデータに適用した。その結果、DNA コピー数と mRNA 発現量の間の相関が高い遺伝子クラスター、およびこれに関連するモジュールのクラスターを発見し、それは既存の分子生物学的な研究結果とも合致する結果であった。さらに、この結果をもとに生存時間分析を行ったところ、免疫系に関わるモジュールの活性の程度が大腸がん患者の生存率に有意に影響を与えていることがわかった。この結果から、提案手法はがん患者の生存期間を予測するための予後因子を同定する手法としても期待できる可能性があることがわかった。

- 主な研究成果:

  バイオインフォマティクス分野の査読付き国際会議 BICoB に論文が採録された。

  1. 査読付き国際会議:

     (a) Yoichi Chikahara, Atsushi Niida, Rui Yamaguchi, Seiya Imoto, Satoru Miyano, "Integrative Clustering of Cancer Genome Data using Infinite Relational Models", Proc. of the International Conference on Bioinformatics and Computational Biology Honolulu, Hawaii, USA, Mar. 2015 (BICoB2015)

3　可変ステップ幅数値積分を用いた生化学反応シミュレーション

- **研究目的:**
  遺伝子制御、シグナル伝達、代謝といった複雑な生化学反応の挙動を理解するために、分子の移流を、反応速度に関する常微分方程式 (ODE) などの数理モデルで表し、分子濃度に関する数値計算シミュレーションを行う研究が盛んに行われている。

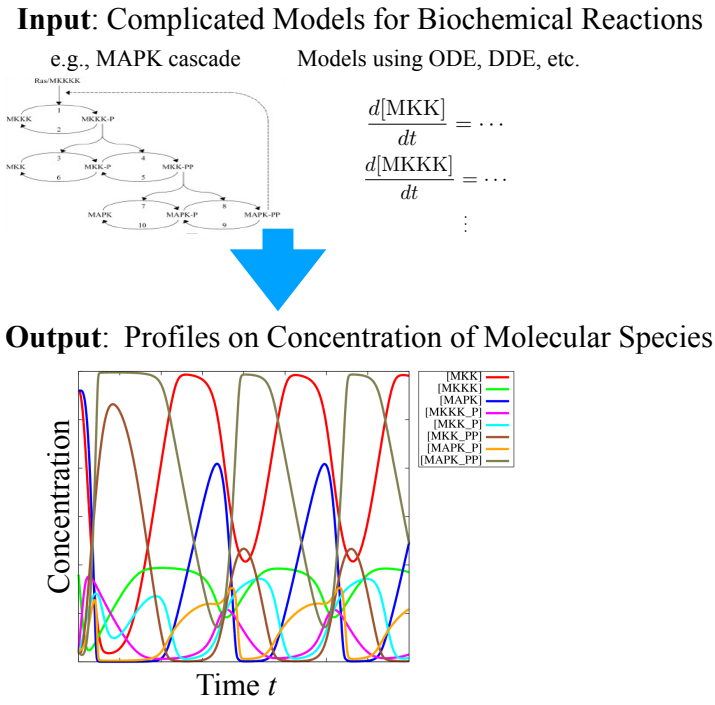

図4　微分方程式モデルに基づく分子濃度シミュレーション

生化学反応を数理モデルとして記述するための言語規格として SBML(Systems Biology Markup Language) があるが、複雑かつ多様な生化学反応を記述するためにその仕様は極めて膨大で、全ての仕様に準拠したシミュレータは先行研究で開発された LiBSBMLSim のみであった。LiBSBMLSim は、与えられた ODE の数値解を得るために値が固定されたステップ幅に基づいて数値積分を行うが、このような固定ステップ幅による数値積分アルゴリズムでは、無限小のステップ幅に固定すれば正確な数値解を得られるものの、それは計算効率の観点から非現実的であり、またユーザが数値解の精度と計算効率のトレードオフに基づいて適切にステップ幅の値を定めるのは一般に難しいという問題があった。
そこで本研究課題では、ユーザが指定する数値解の精度を保証しながらステップ幅を調節するアルゴリズムとして可変ステップ幅数値積分アルゴリズムを LiBSBMLSim に導入するとともに、同アルゴリズムを用いた場合の遅延微分方程式モデルの数値解計算など、SBML の複雑な仕様に対応するための方法を考案することで、SBML の全ての仕様に準拠した計算効率の良い生化学反応シミュレーションを実現することを目指す。

- **技術的背景:**

  ODE と初期条件を与えたもとで関数値を求める問題を ODE の初期値問題という。一般に、ODE の初期値問題は積分が解析的に行えないケースが多く、それゆえ数値解法に頼らざるを得ないケースが多い。ODE の初期値問題を数値的に解くアルゴリズムは固定ステップ幅数値積分と可変ステップ幅数値積分の 2 つに大別される。前者は指定したステップ幅を十分微小な区間とみなし、関数の傾きを Taylor 近似によって近似し、この傾きの値を用いて次の時刻での関数値を求める。これに対し後者では、同様の方法で Taylor 展開の p 次近似解と $(p+1)$ 次近似解を求め、それらの差から離散化誤差を推定し、その推定値がユーザの指定する許容誤差を満たすよう、ステップ幅を調節しながら各時刻での関数値を求める。ここで、一般に離散化誤差は関数の勾配が大きいほど大きくなるため、勾配が大きい区間では小さなステップ幅の値が採用され、勾配が小さい区間では大きなステップ幅の値が採用されることになる。

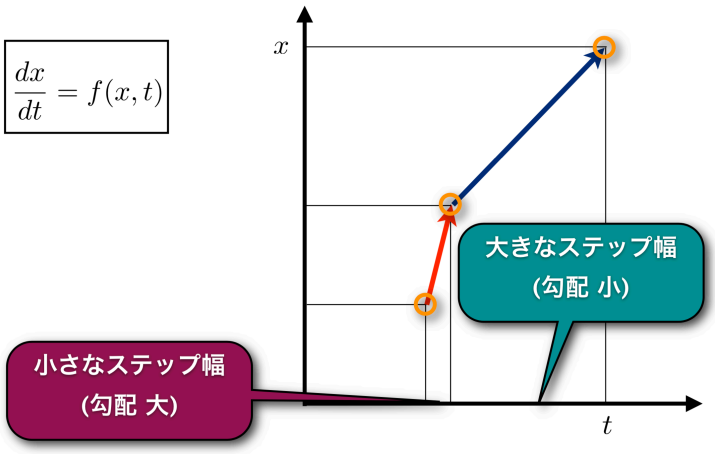

図5　可変ステップ幅数値積分におけるステップ幅の調節

本研究では、可変ステップ幅数値積分アルゴリズムを用いて、数値精度と計算効率のトレードオフを考慮しながら分子濃度に関する数値計算シミュレーションを行うことを目指す。

- **研究の概要:**

  本研究では、既存の可変ステップ幅数値積分アルゴリズムである Runge-Kutta-Fehlberg 法、および Cash-Karp 法を数値計算シミュレータ LiBSBMLSim に導入し、同アルゴリズムを利用したもとでも SBML の全ての仕様を充足することを示した。特に SBML の仕様の一つである遅延微分方程式モデルを数値的に解くためには、指定された遅延に基づいて過去の関数値を評価する必要があるが、可変ステップ幅数値積分の場合、近傍の時刻における関数値が求められていない場合が多く、これを評価することができない。本研究では遅延に即した過去の関数値を線形補間で近似的に評価することで、指定された許容誤差の範囲内で遅延微分方程式を解くことを可能にした。

  簡単な 3 変数の ODE モデルを用いて、計算速度を固定ステップ幅数値積分アルゴリズムである古典的 Runge-Kutta 法と比較したところ、Cash-Karp 法は計算に要するステップ数が 1/129 倍で、CPU 時間は 1/78 倍であり、十分高速化が達成できていることが確認できた。

- 主な研究成果:

  バイオインフォマティクス分野の最難関国際論文誌 Bioinformatics に共著論文が採録された。

  1. 査読付き国際論文誌 (共著)

     (a) Hiromu Takizawa, Kazushige Nakamura, Akito Tabira, Yoichi Chikahara, Tatsuhiro Matsui, Noriko Hiroi and Akira Funahashi; "LibSBMLSim: A reference implementation of fully functional SBML simulator"; Bioinformatics 29.11 (2013): 1474-1476.