因果性発見を用いた時系列データからの情報の構造化・モデル化手法

渋谷 崇† 原田 達也† 國吉 康夫†
†東京大学大学院情報理工学系研究科

1 はじめに

近年, データを蓄えるデータベースの技術も向上している中, 蓄えられた大量のデータから知識を獲得しようとするデータマイニングの研究が盛んになっている [1]. データからの知識獲得方法の一つに情報の構造化・モデル化がある. これは様々なデータからそれらの間にある因果的な関係性をグラフ化し, データの予測などに利用しようとするものである. 一方, 世の中には, ものの有無や天気といったシンボルで得られる情報と, 経済指標や気温といった連続値で得られる情報とがある. 多変数の連続値時系列データから因果性を用いて構造化・モデル化する手法は Tikka らの手法 [2] など多く提案されているが, これらの手法はシンボルデータと連続値データを統一的に扱えない.

そこで本研究では, 時系列のシンボルデータと連続値データを統一的に扱える 情報の構造化・モデル化手法を提案する.

2 提案手法

2.1 因果性定量化手法

因果指標と呼ばれる, 2 変数間の因果性を定量化する指標がある. 因果とは原因と結果の関係なので, 2 変数に関して非対称な指標となる. 因果指標の 1 つに Schreiber によって提案された Transfer Entropy [3] がある. これは一方の変数がもう一方の変数に与えている情報量を算出することで因果性の強さを計るという, 情報理論に基づいた指標である.

X, Y は時刻 t において x_t, y_t というデータをもつ時系列の変数とする. これらの時系列データはそれぞれオーダー k, l の定常マルコフ過程と近似できると仮定する. このとき, X から Y への Transfer Entropy は以下の式で与えられる.

$$T_{X\to Y} = \sum_{y_t,\, \boldsymbol{y}_{t-1}^{(l)},\, \boldsymbol{x}_{t-1}^{(k)}} p(y_t, \boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)}) \log_2 \frac{p(y_t \mid \boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)})}{p(y_t \mid \boldsymbol{y}_{t-1}^{(l)})}. \tag{1}$$

ここで, $\boldsymbol{x}_{t-1}^{(k)}$ はベクトル $(x_{t-1}, x_{t-2}, \ldots, x_{t-k})$ を意味し, 埋め込みベクトルと呼ばれる. この指標は情報理論に基づいた指標で, シンボルデータに対してはデータに処理を加えることなく適用できるが, 連続値データに適用する場合は連続値データを量子化しなければならない. この場合, 量子化の bin 数がパラメータとして必要になってしまう.

そこで, 本研究では連続値データが自己回帰過程に従うと仮定することで導出される, 連続値データとシンボルデータをパラメータに依らず, シームレスに扱うことのできる因果指標を提案する.

まず, 連続値データ X から連続値データ Y への因果性を定量化する指標は以下の式で与えられる.

$$T_{X\to Y} = \frac{1}{2} \log_2 \frac{\sigma_y^2}{\sigma_{y|x}^2}. \tag{2}$$

ここで, σ_y^2 は自己回帰モデル $y_t = \boldsymbol{a}^T \boldsymbol{y}_{t-1}^{(l)} + \epsilon_t^{(y)}$ における誤差 $\epsilon_t^{(y)}$ の分散, $\sigma_{y|x}^2$ は混合回帰モデル $y_t = \boldsymbol{b}^T \boldsymbol{y}_{t-1}^{(l)} + \boldsymbol{c}^T \boldsymbol{x}_{t-1}^{(k)} + \epsilon_t^{(y|x)}$ における誤差 $\epsilon_t^{(y|x)}$ の分散である.

次に, シンボルデータ X から連続値データ Y への因果性を定量化する指標は以下の式で与えられる.

$$T_{X\to Y} = \frac{1}{2} \sum_{\boldsymbol{x}_{t-1}^{(k)}} p(\boldsymbol{x}_{t-1}^{(k)}) \log_2 \frac{\sigma_y^2}{\sigma_{y|x}^2\left(\boldsymbol{x}_{t-1}^{(k)}\right)}. \tag{3}$$

ここで, $\sigma_{y|x}^2\left(\boldsymbol{x}_{t-1}^{(k)}\right)$ は Switching 自己回帰モデル $y_t = \boldsymbol{b}\left(\boldsymbol{x}_{t-1}^{(k)}\right)^T \boldsymbol{y}_{t-1}^{(l)} + \epsilon_t^{(y|x)}$ のシンボル $\boldsymbol{x}_{t-1}^{(k)}$ に対応する誤差 $\epsilon_t^{(y|x)}$ の分散である.

そして, 連続値データ X からシンボルデータ Y への因果性を定量化する指標は以下の式で与えられる.

$$T_{X\to Y} = \frac{1}{2} \sum_{y_t,\, \boldsymbol{y}_{t-1}^{(l)}} p(y_t, \boldsymbol{y}_{t-1}^{(l)}) \log_2 \frac{\left|C_{\boldsymbol{x}_{t-1}^{(k)}}\left(\boldsymbol{y}_{t-1}^{(l)}\right)\right|}{\left|C_{\boldsymbol{x}_{t-1}^{(k)}}\left(y_t, \boldsymbol{y}_{t-1}^{(l)}\right)\right|}. \tag{4}$$

ここで, $\left|C_{\boldsymbol{x}_{t-1}^{(k)}}\left(\boldsymbol{y}_{t-1}^{(l)}\right)\right|$ はシンボル $\boldsymbol{y}_{t-1}^{(l)}$ に対応する $\boldsymbol{x}_{t-1}^{(k)}$ の分散共分散行列の行列式である.

式 (1)~ 式 (4) は全て, 情報理論に基づいた指標で, X が Y に与えている平均ビット数を表している. 単位を bit に揃えることによって, シンボルデータと連続値データをシームレスに扱うことが可能になった.

2.2 因果性を用いた情報の構造化・モデル化手法

まず, 前節で提案した因果指標を用いて情報を構造化する手法を提案する. いま N 個の変数についてデータが得られたとすると, 自身から自身への因果指標値

Structuring and modeling method of information from time-series data using causality detection
†T. SHIBUYA, †T. HARADA, and †Y. KUNIYOSHI
†The University of Tokyo

を除いて $N(N-1)$ 個の因果指標が算出できる．その $N(N-1)$ 個の因果指標に対して閾値を導入し，重要な因果性とそうでない因果性を分けることで構造化することができる (図 1)．

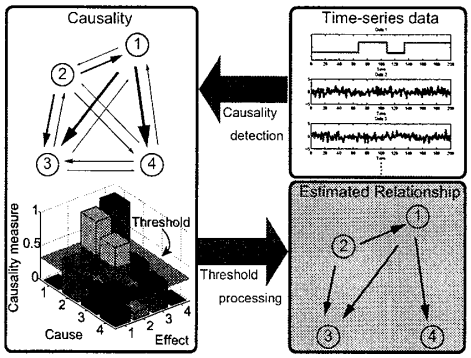


図 1: Overview of proposed method.

モデル化は因果指標値を重みとして，因果指標の算出に用いた遷移確率分布を組み合わせることで行う．ここでは次章の実験で取り扱う連続値データのモデル化手法を示す．いま N 個の変数のうち，モデル化を行う連続値変数を Y とし，それ以外の変数を $X_1, \ldots, X_{N-1}$ とする．このとき因果指標 $T_{X_i \to Y}$ を算出するのに用いたモデル $y_t = f(\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{i,\,t-1}^{(k)})$ を組み合わせて，以下の式でモデル化が可能である．

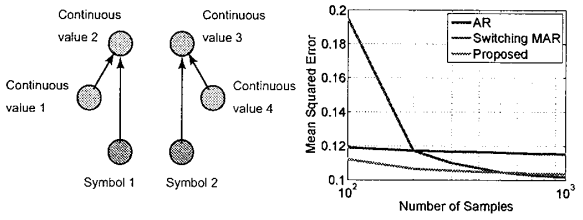
$$y_t = \frac{\sum_{i=1}^{N-1} T_{X_i \to Y} f(\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{i,\,t-1}^{(k)})}{\sum_{i=1}^{N-1} T_{X_i \to Y}} \tag{5}$$

$$f(\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{i,\,t-1}^{(k)}) = \begin{cases} \boldsymbol{b}_i^T \boldsymbol{y}_{t-1}^{(l)} + \boldsymbol{c}_i^T \boldsymbol{x}_{i,\,t-1}^{(k)} & X_i \text{ が連続値の場合} \\ \boldsymbol{b}_i \left(\boldsymbol{x}_{i,\,t-1}^{(k)}\right)^T \boldsymbol{y}_{t-1}^{(l)} & X_i \text{ がシンボルの場合．} \end{cases} \tag{6}$$

3 実験

前節で提案したモデル化手法の性能を調べる実験を行った．

実験は図 2 (a) のような因果的関係性を持つ 6 変数からなる人工の時系列データを用いて行った．6 変数のうち，4 変数が連続値データ，2 変数がシンボルデータであり，これらは全て単純マルコフ過程に従う．この因果的関係性に基づいて異なる初期値から生成した 100 の時系列データセットに対して，交差検定を行って手法ごとの予測精度を比較した．ここでの予測精度とは，ある時間ステップでの各変数の値が与えられた時に，次のステップの値を学習したモデルによってどれだけ正確に予測できるかをさす．連続値データの予測精度について，比較した手法は 3 つである．因果性を無視した各変数での自己回帰，2 つのシンボルデータが切り替わるごとに多変量自己回帰モデルが切り替わる Switching 多変量自己回帰，そして提案手法である．結果を図 2 (b) に示す．



(a) Causal relationship of data. (b) Comparison of methods.

図 2: Experiment with artificial data.

図 2 (b) はデータのサンプル数に対して予測値の平均 2 乗誤差をプロットしたものである．この結果から，提案手法は因果性を踏まえたモデル化を行っているため，通常の自己回帰より精度が高いことがわかる．また，Switching 多変量自己回帰との比較では，提案手法はサンプル数が少ない場合に優位な性能を示している．これは，多変量自己回帰ではモデル化に必要なサンプル数が変数の数に応じて増えるが，提案手法は 2 変数間のモデルの集合であるため，サンプル数が少なくても比較的正確な予測が可能だと考えられる．

同様の実験を Yu らの遺伝子調整シミュレーションデータ [4] でも行った．このシミュレーションデータは変数の数が 12 以上あり (全て連続値)，より複雑な因果的関係性を持っている．各変数での自己回帰，多変量自己回帰，提案手法の 3 つの手法について比較実験を行ったところ，図 2 (b) と同様の結果が得られた．

以上の実験から，提案手法はデータのサンプル数が少ない場合でも因果性を踏まえた，精度の高い予測が可能であると言える．

4 まとめ

本研究では，シンボルデータと連続値データをシームレスに扱うことができる因果指標を提案し，さらに因果指標をもとに情報を構造化・モデル化する手法を提案した．また，人工データを用いた実験によって，サンプル数が少ない場合でも因果性を踏まえたモデル化が可能であることを示した．

参考文献

[1] Q. Luo, Advancing knowledge discovery and data mining, *WKDD 2008*, pp. 3-5, 2008.

[2] J. Tikka and J. Hollmen, Learning linear dependency trees from multiple time-series data, In *Proc. of the ICDM2004 Workshop on Temporal Data Mining*, 2004.

[3] T. Schreiber, Measuring information transfer, *Physical Review letters*, Vol. 85, No. 2, pp. 461-464, 2000.

[4] J. Yu, V. A. Smith, P. P. Wang, A. J. Harteminkd and E. D. Jarvis, Advances to bayesian network inference for generating causal networks from observational biological data, *Bioinformatics*, Vol. 20, No. 18, pp. 3594-3603, 2004.