知識獲得の計算理論に向けて：多変量情報流による潜在的機構の推定

Toward Computational Theory of Knowledge Acquisition: Estimating Latent Mechanisms Through Multivariate Information Flows

日高　昇平†
Shohei Hidaka

†北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
shhidaka@jaist.ac.jp

Abstract

科学的活動の多くにおいて、現象の観察・経験を通じ、直観に沿う仮説を立て、それに基づく理論構築が行われる。このような理論構築過程の一つの定式化として、ある多変量時系列を付与のデータとし、それを生成する潜在的な概念モデルの推定問題を考える。本研究では、情報理論を一般化した多変量双方向情報理論を用いる事で、現象に関する事前知識性を必要とせずに、データのみから一般的な潜在的な概念モデルを推定できる事を示す。

Keywords — Knowledge acquisition, Information theory, Multivariate time series

1. 知識獲得の暗黙的計算過程

科学的活動の一つの目的は、目的観察・観測された現象そのもの(データ)から、それを生成・近似するより簡潔な表現形式(理論・モデル)を構築する事である(図 1)。もし構築されたモデルが、何らかの意味で一般性をもち、有意義であるならば、それは多くのコミュニティ(研究グループ・学会・社会など)のメンバーに共有される知識となりえる。このようなモデル化・理論化の過程では、多くの場合、現象・データの観察・経験(暗黙知)を通じ、観察者の“直感”に沿った仮説を立てることから始められる。構築されたモデルあるいは理論は形式的であるのに対し、それを生成する観察者の直感は暗黙的である。従って、知識の獲得過程を理解するうえで、この暗黙的計算過程の解明が重要である。

本研究では、データから知識を構築する際の鍵となる「観察者の暗黙知」の一つのモデルとなりうる理論体系を提案する。具体的な問題として、観察データがある多変量時系列として与えられた場合に、それを生成する可能性の高いモデルのクラスを特徴づける問題を考える。これは、任意の形式の N 変量時系列から、その N 変量の間の確率的・力学的な 2 項関係を N^2 の方向つき情報流として定量化する問題として定式化できる(図 2a)。ここで得られる情報流ネットワークは、観察された要因間の関連性の度合いに関する直感、あるいは「前モデル的」な概念グラフとみなせる。

このような前モデル的なデータの記述を、ある知識獲得過程における「観察者の暗黙知」として考えるには、少なくとも以下の２つの要件を満たすべきである。

（１）少事前知識性：観察事象に対する最小限の事前知識で前モデル的仮説を構成可能である。

（２）汎用性：特定の機構の生成するデータだけでなく、多様なデータクラスに適用可能である。

この２つの条件に対し、本研究では、（１）与えられた時系列データの観測変数の一貫性・同期性のみの前提に基づき、（２）多様な確率的・力学的なデータ生成過程に対し、その潜在する情報的メカニズムを推定可能な理論的枠組みを示

す。具体的には、情報理論(Shannon, 1948)あるいはその拡張である双方向情報理論(Marko, 1973; Schreiber, 2000)を、さらに多変量へと一般化した双方向情報理論(Hidaka, 2012)を提案する。

多変量双方向情報理論の計算過程は、典型的な科学的方法論を抽象化したものとみなせる。多くの場合、科学者はある２つの因子を、他の因子の影響をなるべく統制・固定した上で調べ、その２因子間の関係性を見極める。これと同様に、多変量双方向情報理論では、ある変数 X から変数 Y への情報の流れを、その他の変数 Z の影響を差し引いた上で(条件つきで)推定する(Hidaka, 2012)。一方、(二変量)双方向情報理論(Marko, 1973)では、第三の変数 Z が存在していても、その影響を統制せずに、ある変数 X から変数 Y への情報の流れを(Z の影響を含みながら)推定する。従って、多変量双方向情報理論によって、単に情報量の算出というよりも、複数変数間の関係性を考慮した上でのデータの持つ情報的特性を定量化する事ができる。本研究では、知的営みにおいて重要な過程の一つである概念モデルの構築は、こうした変数間の関係性を考慮した上での情報量の推定とみなせると仮説をたて、これを検討した。以下では、シミュレーション研究および実データの分析研究について説明する。

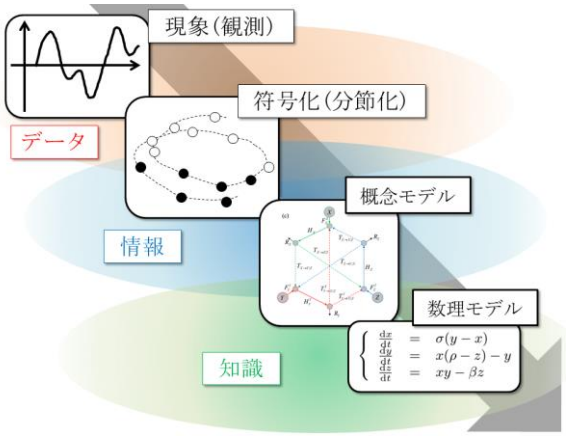

図 1: 知識獲得過程の概念モデル

2. 知識獲得過程のシミュレーション：力学系データから概念モデルの同定

多変量情報理論による関係グラフの推定のケーススタディとして、まず因果関係を操作することの出来る理論的な対象に対し、その潜在的な因果関係が未知な場合に双方向多変量情報量のみでどの程度正確に関係グラフが推定可能であるかシミュレーションを行った。因果関係が決定論的に決まり、しかし、長期的な予測が困難な複雑な対象として、非線形力学系が挙げられる。非線形力学系では、ある時点から次の時点への状態変化が、ある非線形方程式(連続系では微分方程式)によって与えられる一方、多くの場合、その挙動は複雑でカオスと呼ばれる。このシミュレーションでは、ある複数変数のカオス的挙動の系列を観測した場合に、その経験的な時系列から、系に固有の変数間の関係を情報の流れとして記述する。これは、まさに複雑な多変量関係を観測したときに、その変数間の関係性を特徴づける概念モデルの構築に対応する。シミュレーションでは、“正しい” 概念モデルを非線形力学系のパラメータとして操作し、多変量情報理論によってその潜在的な関係性をどの程度正確に推定できるか評価した。具体的な系として、結合写像格子(Coupled Map Lattice: CML)を用いた。CML は比較的単純な 1 次元力学系を結合し、個々の要素の関係性を任意に操作可能であり、かつ個々の要素だけでは起きない創発的な特性を全体として示す(Kaneko, 1992)。

表 1：CML 多変量データに基づく関係グラフの推定結果

問題設定				正解率 多変量		正解率 二変量	
変数の数	関係対組合わせ	ケース数	関係対の数	ケースベース	変数対ベース	ケースベース	変数対ベース
3	2^6	16	96	100%	100%	75.0%	93.75%
4	2^{12}	218	2,616	90.78%	97.78%	9.22%	70.67%
5	2^{20}	9,608	192,160	81.48%	95.41%	1.24%	71.23%

表 1 は、変数の数を 4,5 と増やした場合に、全ての組み合わせ(それぞれ 218, 9608 ケース)の変数対ごと、そしてケースごとの正解率を示している。ケースごとの正答率は個々のケースで 1 つでも推定誤りが含まれる場合は不正解としているため、変数対毎の正答率より厳しい基準となっている。この結果は、多変量移送情報量はケースごとで 90.78%(4 変数), 81.48%(5 変数)、変数対ベースで 97.78%(4 変数), 95.41%(5 変数)と比較的高い推定精度を保つ事を示している。一方、比較基準となる二変量移送情報量はケースごとで 9.22%(4 変数), 1.24%(5 変数)、変数対ベースで 70.67%(4 変数), 71.23%(5 変数)と、変数が増えると急激に推定精度が下がる事を示している。この結果は、変数の増加に対し、多変量移送情報量は比較的高い推定精度を保つ事を示し、より高次な依存関係に対する堅牢性を示唆している。これに対し、二変量移送情報量がこの堅牢性を持たない事、また両者の理論的性質の違いを考慮すると、関係グラフの推定において、(1) 情報量の推定に加えて (2) その高次の従属関係(交絡関係)を明確に分離する(条件わけする事)の 2 点が要点として挙げられる。

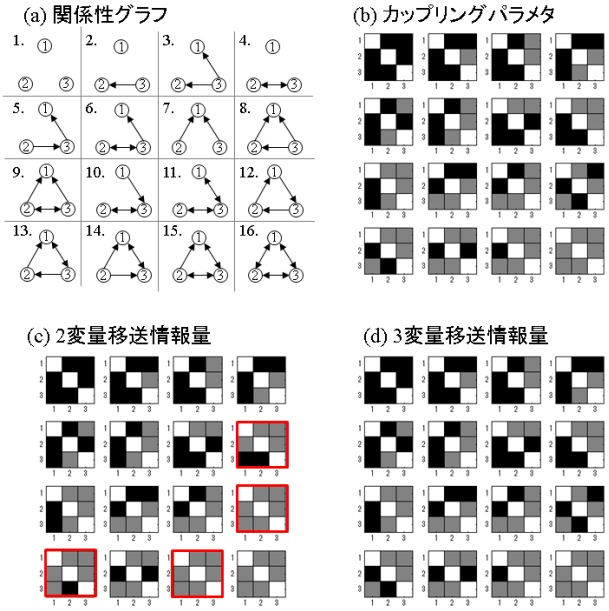

図 2: (a) 関係性グラフ(3 変数の場合、全 16 組み合わせ), (b) 結合写像格子の結合パラメータ, (c) 二変量移送情報量の推定値(非対角要素), (d) 三変量移送情報量の推定値(非対角要素)

3. 実データへの応用：身体運動の情報流ネットワーク

次に、実際の経験的なデータに対する多変量双方向情報理論の適用を行い、概念モデルの形成過程を検討する。複雑な要因間の相互作用のある対象の一つとして、身体運動の分析を行った。身体運動は環境、身体、神経活動の三者の間の複雑な相互作用によって成立する。複雑な多関節運動の場合、身体各部はそれぞれに協調しながら一つの運動を生成すると考えられている(Yamamoto, Ishikawa, & Fujinami, 2006)。ある一つの運動の習得は、必ずしも身体部位間の静的な関係性の獲得を意味しない。むしろ、熟達に伴って、運動はより効率的に、より効果的に変化すると考えられる。こうした熟達した運動の獲得をスキル学習と呼ぶ(Leonard, 2002)。しかし、スキルと呼ばれる複雑な運動に関して、身体運動の熟達をどのように捉えるべきかは個別の課題に特化してケーススタディとして行われており、未だに統一的な理論的解釈は与えられていない。たとえば、球技(e.g., バスケットボール)

のように、明確に運動の標的が定義されている特定の運動(ゴールにボール投げ込む)の場合、その成功率として習熟を一つの尺度で表す事が可能である。しかし、こうした明確な標的が定まらない、たとえば演劇、歌唱や音楽演奏、描画など運動においても、人はその運動の質(あるいはその結果としてのパフォーマンス)をある程度一貫して評価する事が出来る。こうした背景を踏まえ、複数の身体部位間の協調関係を反映する概念モデルをデータから推定し、熟達化に伴う概念モデルの変化を検討する。

3.1 分析手続き

対象としたのは、サンバの演奏運動中の身体動作をキャプチャしたデータである(Yamamoto, Ishikawa, & Fujinami, 2006)。この実験では全身の 16 の特徴点と楽器(シェイカー)の両端 2 箇所の計 18 箇所にマーカーをつけた演奏家 5 名が、5 つのテンポ(60, 75, 90, 105, 120 拍毎分(BPM))で演奏を数分間を行った。こうして得られたデータのうち、本研究では、特に熟達に顕著な差があると見られる 3 名(40 年以上の経験を持つ上級者, 5 年の経験をもつ上・中級者、2 年の経験をもつ中級者)の演奏に直接関わる右腕(肘・手首)と楽器(両端 2 箇所)の 4 箇所の動きを分析した。各演奏者につき、4 箇所の身体動作は 3 次元の空間上の点列として 86.1Hz のレートでサンプルされたが、計測に伴う高周波ノイズのため、その半分のデータ点(43.05Hz に相当)を分析に用いられた。各時系列データは、演奏者、演奏リズムごとに記号的最近傍法(Buhl & Kennel, 2005)を用いて、符号化をした後、4 次マルコフ性を仮定した上で、その遷移確率分布の N 変量移送情報量を算出した。前述したシミュレーションの分析と同様に、推定された移送情報量を 0 とする帰無仮説に対し統計検定を行うことで、$\alpha = 0.01$ 水準で有意に 0 より大きい移送情報量を求め、これを関係グラフにおける情報の流れとして定義した。

3.2 結果・考察

経験年数の違いから期待される熟達の度合いに応じて、上級者、中上級者、中級者とした。リズム運動の多変量時系列(楽器 1, 2, 手首、肘の 4 箇所)から推定された情報流グラフを図 4 に示す。それぞれの演奏家に対し、5 つの演奏テンポ条件、12 関係対について、有意に 0 より大きい情報流の割合 (12 ペア・5 条件、のべ 60 ペア中)をバーグラフに、その条件間の標準偏差をエラーバーとして表示している。熟達の度合いの高いほどに、演奏に関わる右腕および楽器に間における有意な情報流の割合が高い事が分かった。また、バーグラフの上部には、それぞれの演奏家で 5 条件すべてで有意に 0 より大きい身体部位間の関係を方向つきの関係グラフとして示している。推定された関係グラフは、上級者ではほとんど全ての身体部位間で情報の流れがあるのに対し、中級者は手首と楽器の一部のみに有意な情報がある事を意味している。また、中上級者では、中級者と上級者と中間程度の情報の流れが見られた。この中上級者の関係グラフでは、肘から手首、手首から楽器への情報の流れがあり、これは腕のしなりによるむち運動で楽器に動きが伝達する状態を表していると考えられる。これに加えて上級者では楽器から手首・肘への情報の流れも見られ、楽器の状況によって手首・肘の運動を変化させていると解釈できる。以上の結果は、上級者ほど、全ての関連する身体・楽器の間で双方向に情報伝達が起こり、腕から楽器に運動が伝わるだけではなく、楽器から腕への情報がその制御に重要である事が示された。

こうした結果は、予想された通り、熟達によって身体運動が巧妙になる事を示すだけでなく、多変量情報理論によってその具体的な関係性を構築できた事を意味する。本実験で調べた複雑な動きの精妙な制御は、未だに解明すべき部分が多い。数十ミリ秒の精度で多数の間接・筋肉が制御され、作り出された一つの洗練された動きに、我々はそれをどのように、熟達を感じ取

るのだろうか。本分析は、このような問いに対して、一つの可能性を示している。つまり、我々が他者の身体運動の観察を通じて、情報的な関係性を推定し、その情報流ネットワーク上の密度として、運動の精妙さを感じ取るのではないだろうか。本研究は、ひとつの事例研究に過ぎないが、多変量双方向情報理論はこうした複数の要因間の複雑な時間的なパタンから、それを整理する有用な関係グラフを構築できることを示している。

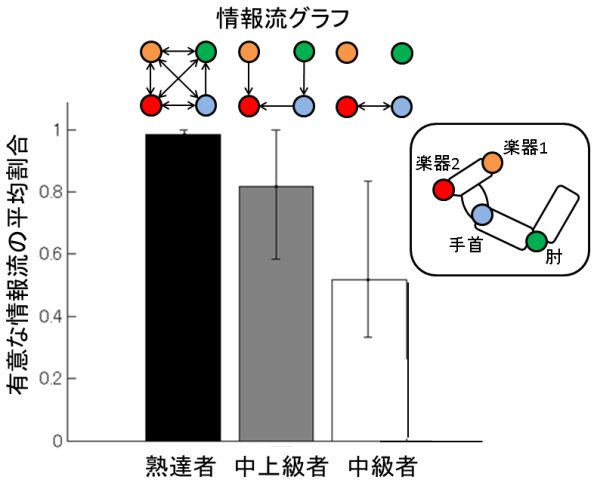

図 4: リズム運動の多変量時系列(楽器 1, 2, 手首、肘の４箇所)の情報流グラフ

4. 総合討議

本研究では、データから知識を獲得する過程において、最も初期に行われる概念モデルの形成について理論的検討を行った。概念モデルでは、ある現象に関してそれを適切に代表するいくつかの状態の分節化、そしてそれらの間の関係性の記述を行う。この概念モデルはそれに続くより詳細な数理モデルなどの基礎となり、あるいは社会科学など数理的なモデル化が困難な複雑な現象の説明においては中心的な役割を果たす。一方、概念モデルの形成過程そのものは、データ観察者の暗黙知であり、最終的な表現である関係グラフを除いてその内省的メカニズムは不明な点が多い。本研究では、多変量時系列データの観察から、その変量間の情報論的な関係グラフを推定する過程と概念モデルの構築とみなし、これを最小限の事前知識により行う理論体系を提案した。

本研究で検討した多変量双方向情報理論(MTE)は、Shannon の(一方向)情報理論、Marko の二変量双方向情報理論を拡張した一般化理論である(Hidaka, 2013)。MTE では、複数の変量間の関係(高次情報量)を方向つきの情報の流れに分解して表現する。MTE の計算過程は、典型的な科学的な思考法と同様に、他の交絡因子を統制した上で、二変数 X と Y の関係を特定する事を情報理論の拡張として表現したとみなせる。

理論モデルを用いてデータを生成したシミュレーションでは、第三の交絡因子の関与を考慮しない PTE よりも、MTE を用いた場合に高精度で関係グラフの構築が可能である事が示された。これに続く身体運動データの分析では、熟達者の運動と中級者の運動の違いが、身体部位間の依存関係の密度という形で、定量化できることが明らかになった。分析では、一切の身体部位間に関する事前知識を与えず、データだけから具体的な関係グラフの図式化したにも関わらず、熟達者の運動がより精密かつ複雑な構造を持つ、という我々の直観に合致した関係グラフが得られた。こうした一連の結果は、MTE による関係グラフ構築が、概念モデルの構築に類似した性質を持っていることを示唆している。これは、序論で提示した少ない事前知識の下で関係グラフを構築という条件（１）を満たすことを意味する。さらに、3,4,5 変数の全ての可能な関係グラフの組み合わせを分析したシミュレーションの結果と、さらにノイズなど必ずしも理想的ではない条件下で得られた身体運動データの分析結果を考慮すると、MTE は少なくとも一定程度の汎用性（条件（２）)を有すると考えられる。Hidaka (2012)は、本研究で検討した２つのケースに加えて、さらに、理論力学系の一つである Lorenz 系への適用、また生理学的データ分析への応用も示している。この二条件については、今後も MTE による分析を理論的・経験的な側面から検証を重ねる事で、どの程度の

事前知識が必要であるか、また汎用性を持つかを調べていく必要がある。

なぜ MTE による情報流ネットワークは、その背景にある(数理的)モデルを推定せずに、変数の間の従属性を定量化できるのだろうか。通常はある数理的モデルを立てた上で、そのモデルの当てはめを行った上で、関係性を特定する事を考えれば、潜在する数理メカニズムの特定前に、その関係性を定量化できるのは不思議ではある。このように数理モデルを特定せずに、変数間の情報量を特定できるのは、情報理論の持つモデルフリー性にある。情報理論において、モデルはある現象の符号化に関わる。ある空間を十分に高い精度で符号化を行った場合に、どのような座標系で符号化しても情報量が不変になる場合に、モデルフリーと言う。多変量情報理論の特殊な場合である二変量移送情報量に関して、このモデルフリー性が成り立つ事が示されている(Kaiser & Schreiber, 2002)。この理論的性質により、ある現象をどのように符号化するか、という詳細に立ち入らずに、多変量双方向情報理論は対象の従属性を定量化することができる。つまり、少事前知識性・汎用性を同時に満たしながら、データから概念モデル的な構造を推定しており、これは概念モデル形成過程を説明するモデルになりうる。

最後に、本研究で提案する多変量情報流で、現状では扱えない対象についての考察をについて述べる。情報理論の拡張たる多変量情報流は、情報理論と同じく、符号化されたデータを対象に情報量を測る。逆に言えば、情報理論および多変量情報流が、推定すべき潜在的メカニズムが不明な対象に適用される場合、データがどのように符号化されるべきか、符号化の良さについては、言及することができない[1]。したがって、情報理論の枠組みの外に、対象そのものの符号化についての理論が必要となる。特に本研究で対象とした非線形力学系に関しては、生成分割と呼ばれる“最も良い符号化”の存在が知られている場合があり、時系列データからの推定法もいくつか提案されている(Hidaka & Yu, 2010a; 2010b)。こうした汎用的な符号化の理論と組み合わせることで、多変量情報理論の応用範囲を拡張していくことが今後の課題として挙げられる。

[1] ただし、目標が与えられる場合(教師あり学習・分類問題・情報圧縮など)は、それによって符号化の良さが情報理論によって計算可能である。

謝辞

本稿第 4 節の実データ分析では、藤波努氏の好意により、身体運動のデータを提供していただいた。ここに感謝いたします。本研究の一部は科学研究費補助金(No.23300099)、人工知能研究振興財団、ニューロクリアティブ研究会研究助成の補助を受けた。

参考文献

[1] Buhl, M. and Kennel, M. B. (2005). Statistically relaxing to generating partitions for observed time-series data., Phys. Rev. E 71, 046213.

[2] Hidaka, S., & Yu, C. (2010a). Spatio-Temporal Symbolization of Multidimensional Time Series, International Workshop on Spatial and Spatiotemporal Data Mining, 249-256.

[3] Hidaka, S., & Yu, C. (2010b) Analyzing Multimodal Time Series as Dynamical Systems, 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction, 53-58.

[4] Hidaka, S. (2012). Characterizing Multivariate Information Flows, eprint arXiv:1212.5449

[5] Kaiser, A. & Schreiber, T. (2002). Information transfer in continuous processes. Physica D 166: 42-62.

[6] Kaneko, K. (1992). “Overview of coupled map lattices”. *Chaos An Interdisciplinary Journal of Nonlinear Science*, *2*(3), 279.

[7] Leonard, C. T. (1997). The Neuroscience of Human Movement. Mosby: St. Louis, MO (松村道一，森谷敏夫，小田伸午(訳) ヒトの動きの神経科学).

[8] Marko, H. (1973). “The bidiretional communication theory - a generalization of information theory”. IEEE Transaction on Communication, COM-21(12), 1345-1351.

[9] Schreiber, T. (2000). “Measuring information transfer”. Phys. Rev. Lett., 85(2), 461–464.

[10] Shannon, C. E. (1948). “A mathematical theory of communication”. The Bell System Technical Journal, 27, 379-423, 623-656.

[11] Yamamoto, Y., Ishikawa, K., & Fujinami, T. (2006). Developmental stages of musical skill of samba. Journal of biomechanics, 39, S555.