人工知能学会研究会資料
SIG-AGI-004-03

等価性構造抽出技術の定式化

Formulation of Equivalence Structure Extraction Method

高橋 良暢 [1*]　佐藤 聖也 [2]　山川 宏 [3,4]
Yoshinobu Takahashi[1]　Seiya Satoh[2]　Hiroshi Yamakawa[3,4]

[1] 電気通信大学 情報理工学研究科
[1] Graduate School of Informatics and Engineering, The University of Electro-Communications (UEC)
[2] 産業技術総合研究所 人工知能研究センター
[2] Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST)
[3] （株）ドワンゴ ドワンゴ人工知能研究所
[3] Dwango Artificial Intelligence Laboratory, Dwango Co., ltd.
[4]NPO 法人 全脳アーキテクチャ・イニシアティブ
[4]The Whole Brain Architecture Initiative, a specified non-profit organization

Abstract: Equivalence Structure (ES) extraction method is a technique for extracting a set of K-tuples consist of sequence-ID that can be regarded as equivalent based on similarities that found among sequences in the N-dimensional synchronous sequences. The ES's can be used for analysis of deep neural networks and determination of corresponding markers in imitation learning. In this paper, we provide the definition of ES and the properties of the input data for the ES extraction. In addition, the nature of the ES, and the extraction process are formulated.

1　はじめに

機械学習などに用いられるデータ構造は多くの場合に，属性もしくは特徴量のベクトルの集合を一般化した関係モデルとして与えられるか，その形に変換される．そこで所与のデータから新たな関係モデルを生成できれば，AGI の実現にとって重要な課題である，様々な機械学習におけるより柔軟な推論が可能になると考えられる．

多次元時系列データのような, 系列どうしが互いに時間によって同期可能な構造をもつデータにおいて，系列を識別するための指定子 (系列 ID) のタプルのなかで，等価とみなせるもののペア，またはそれ以上のタプルを含む集合を等価性構造と呼ぶ [1]．等価性構造は画像や動画等の系列データに対する特徴抽出や，通常既知として扱われる見まね学習における教師と生徒の次元対応付け [2] 等種々の応用を考えることができ，現在までに，視覚入力を簡単化した点波シーケンスデータに対して，等価性構造の抽出が可能であることが示されている [1]．関連する研究として，予め決められた特定のパターンがデータ中に出現するかどうかを特定する文字列検索 [3] や，ある次元上に現れるある波形について，それが当次元上に繰り返し現れることを検出するようなモチーフディスカバリー [4] がある．

本稿では，実世界からセンサデータとして多くの場合取得可能な多次元時系列データから，複数の類似する部分的な時系列を取り出して結合することにより，関係モデル [5] に似た，等価性構造と呼ぶ関係特徴の集合を取得する技術とその抽出方法について説明する．具体的にまず 2. 節では入力データとして許される多次元同期系列の性質について述べるほか，そこから抽出することができる等価性構造の定義を記す．3. 節では実際に入力データからどのようにして等価性構造を抽出するのかについて述べる．4. 節では等価性構造抽出技術の現在の研究動向について簡単に述べる．5. 節では今後の課題として，等価性構造抽出技術にどのような応用が考えられるかについて考える．

*連絡先：電気通信大学
〒 182-8585　東京都調布市調布ヶ丘 1-5-1
E-mail:ytakahashi@ni.is.uec.ac.jp

2　各種定義

本節では入力データとなる多次元同期系列と，そこから抽出される等価性構造について定義し，さらに等価性構造が持つ性質について述べる．

2.1　入力データ

N 次元同期系列を考える．ここで N は任意の正整数である．

本稿では入力データの各系列の同期軸として便宜上時間を設定するが，これは時間である必要は特にない．系列を指定する指定子 (以降，系列 ID) として

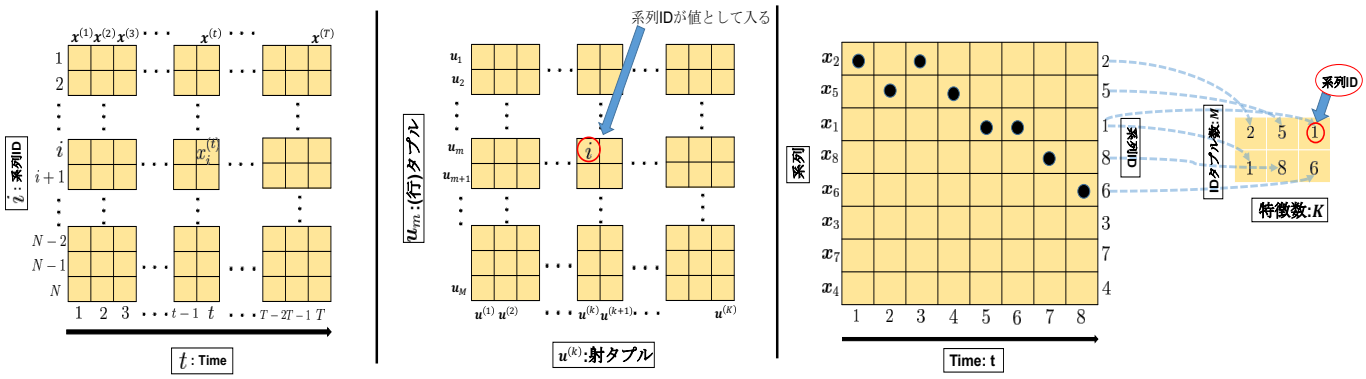

図 1: 左: 入力データとなる多次元同期系列の例，中央: 系列 ID を値とする関係タプルの集合としての等価性構造，右: 8 次元同期系列 (0 または 1 を値としてもつ点波シーケンス) から等価性構造を抽出する例

$i\,(1 \leq i \leq N)$，系列本体を系列 ID によって番号付けし，$\boldsymbol{x}_i \subset \boldsymbol{X}\,(=\{\boldsymbol{x}_i|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_i,\ldots,\boldsymbol{x}_N\})$ とおく．加えて，系列 ID の集合として有限集合 $\boldsymbol{N}=\{i|i\in\mathbb{Z}\wedge i>0\}$ を定義する．$\boldsymbol{N},\boldsymbol{X}$ の要素数 $|\boldsymbol{N}|,|\boldsymbol{X}|$ はともに N である．ここでは時系列を考え，時間 $t\,(1 \leq t \leq T)$ をおき，時間 t における系列の集合を $\boldsymbol{x}^{(t)}$ で表す．以上を図示すると，入力データは図 1 の左の図のように表現することができる．これは多次元時系列において，ある系列内の値は時間 t が決まることでただひとつに決まることを表し，その逆もまた然りであることを表す．

2.2 等価性構造

多次元時系列から抽出する等価性構造は関係モデル [5] に似た性質を持ち，以下のように定義することができる．

定義 1 等価性構造 $\boldsymbol{U}$
N 次元同期系列の各系列 $\boldsymbol{x}_i\,(1 \leq i \leq N)$ に割り当てられる系列 ID: i を値として保持する関係について考える．組として，系列 ID を要素として持つ ID タプルが定義される．これを行タプル，または単にタプルと呼ぶ．本稿では他のタプルとの混同を避けるために行タプルと呼ぶ．行タプルの数は M とし，その指定子を $m\,(1 \leq m \leq M)$，行タプル本体を $\boldsymbol{u}_m$ で定義する．次に列を考える．列数を K とし，各列の要素として M 個全てのタプルから $k\,(1 \leq k \leq K)$ 番目の要素を取得する M-タプルを取得することができる．この M-タプルを射タプル (morphism tuple) と呼ぶこととし，$\boldsymbol{u}^{(k)}$ で表す．ここから $\boldsymbol{u}_m$ は要素数 K の K-tuple($|\boldsymbol{u}_m|=K$) となる．射タプルは M 個の各行タプルのうち，共通している特徴をもつ系列 ID で構成される．以上から系列 ID，即ち値は $u_m^{(k)}$ で表される．以上のようなデータ構造を持つ行タプルの集合を等価性構造と呼び，これを $\boldsymbol{U}=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_M\}$ で定義する．

上記の等価性構造が関係モデルと異なる点として，関係モデルが一般に異なる属性ごとに多様で相異なる属性値が持てることに対して，等価性構造は属性値が系列 ID のみしか存在しないこと，また関係モデルにおける属性が，値が入力される前に属性値が決定されることで何らかの意味付けができることに対し，等価性構造は入力データから値が抽出されて初めて各特徴に意味づけができることの二点である．以上を図にしたものを図 1 の中央に示す．また，多次元時系列データから等価性構造を抽出する簡単な例として，図 1 の右の図に示すような 8 次元時系列点波シーケンスからの抽出例を記す．この例は時間 t に従って 8 系列間を推移する黒点に着目したデータから等価性構造を抽出した例である．

2.3 等価性構造の性質

多次元時系列より等価性構造を抽出する際に，各系列が等価かどうかを判定する際に適切な非類似度関数を用いるが，一般に非類似度関数は等価性構造に対して以下の 2 つの性質をもつ．

2.3.1 行タプル可換性 (tuple commutativity)

図 1 の右の図を例に考えると，$N=8$，$T=8$ の入力データから，系列 ID を要素とする $\boldsymbol{u}_1=(2,5,1)$，$\boldsymbol{u}_2=(1,8,6)$ の二つの行タプルで構成される $M=2$，$K=3$ の等価性構造が抽出されている．任意の等価性構造 $\boldsymbol{U}$ は $\boldsymbol{u}_m$ の集合であるため，要素の順序関係は考慮されない．これは，非類似度関数の以下の性質に起

因する．

$$\begin{aligned} d(\boldsymbol{u}_1, \boldsymbol{u}_2) &= d(\boldsymbol{u}_2, \boldsymbol{u}_1) \quad (1) \\ d((2,5,1),(1,8,6)) &= d((1,8,6),(2,5,1)) \end{aligned}$$

2.3.2 射タプル可換性 (morphism tuple commutativity)

任意の等価性構造 $\boldsymbol{U}$ の列において，k 番目の射タプル $\boldsymbol{u}^{(k)}$ を入れ替えても各行タプル間の非類似度は変わらない射タプル可換性が成立する．

3 等価性構造抽出過程

3.1 入力データ

本稿では入力データとして，図 2 の左に示すような $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8\}$，$N$=8 の 8 次元の系列で構成される，時系列に従う点波シーケンスデータを例として等価性構造の抽出過程を考える．

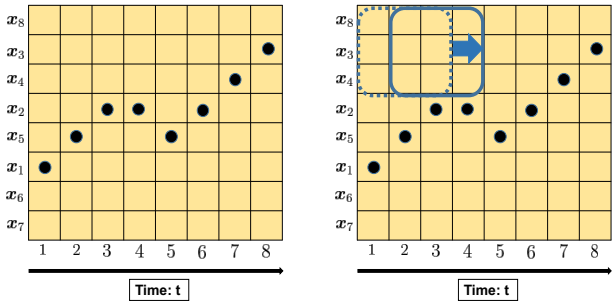

図 2: 左:時系列に沿って流れるように動く黒点，右:入力データを覆う局所シーケンス

3.2 局所シーケンスと LPT(Local Pattern-Tensol)

入力データに対する最初の操作として，図 2 の右の図青枠に示すような，局所シーケンスから LPT(Local Pattern Tensol) を取得する．局所シーケンスは任意の系列 ID 数 K $(1 \leq K \leq N)$ と時間幅 ΔT $(1 \leq \Delta T \leq T)$ で構成されるサイズ $K \times \Delta T$ のウィンドウであり．図 2 の例では K=3，$\Delta T = 3$ となっている．この局所シーケンスを入力データ全てを覆うように 1 時刻ごとに推移させ，局所シーケンス毎に図 3 に示すように，単位時間 τ $(0 \leq \tau \leq \Delta T)$ ごとに局所シーケンスを分解する．次に，系列 $\boldsymbol{x}_i$ における時刻 τ での値が，$x_i^{(\tau)} = \{0, 1\}$ で表されるそれぞれの $\boldsymbol{b}(\tau) = \left(x_1^{(\tau)}, x_2^{(\tau)}, \ldots, x_K^{(\tau)}\right)$，$\boldsymbol{b} = \{\boldsymbol{b}(\tau)\}$ に対して，それぞれの成分が $c_{\boldsymbol{b}(\tau)} = (x_1^{(\tau)} \vee x_2^{(\tau)} \vee \cdots \vee x_K^{(\tau)})$ で表される LPT を生成，格納する．図 3 の例では，$\tau = (0, 1, 2)$，$\boldsymbol{b}(0) = (0, 0, 1)$，$\boldsymbol{b}(1) = (0, 1, 0)$，$\boldsymbol{b}(2) = (0, 1, 0)$ で表される．

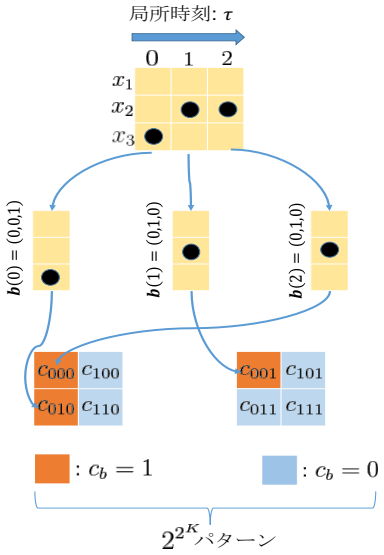

図 3: 局所シーケンスから LPT(Local Pattern Tensol) を生成する様子

3.3 LPT 度数ベクトル

次に局所シーケンスの系列 ID 数 K に依存して変わる LPT 数に対して，入力データ中に発生する LPT それぞれの個数を数え上げる LPT ベクトル $\boldsymbol{F}_p$ $(p\,(1 \leq p \leq {}_N P_K))$ を生成する．ここで p は有限集合 $\boldsymbol{P} = \{p | p = 1, 2, \ldots, {}_N P_K\}$ に含まれるものとする．また，各ベクトルの要素数は $2^{|\boldsymbol{b}(\tau)|(=2^K)}$ であり，各 LPT に対応する要素を指定する指定子として，

$$\begin{aligned} q = c_{000} + 2c_{001} + 2^2 c_{010} + 2^3 c_{011} + 2^4 c_{100} \\ + 2^5 c_{101} + 2^6 c_{110} + 2^7 c_{111} \quad (2) \end{aligned}$$

を定義し，これによって各局所シーケンス毎に $\boldsymbol{F}_p$ 中の要素 $F_{p,q}$ をインクリメントすることで各 LPT の出現頻度を数え．$F_p = \left(F_{p,1}, F_{p,2}, \ldots, F_{p,2^{2^K}}\right)$ を生成する．q の項数は $|\boldsymbol{b}(\tau)|$ に依存して増減することに注意が必要である．

3.4 非類似度計算とクラスタリング

導出したそれぞれの $\boldsymbol{F}_p$，$\boldsymbol{F}_{p'}$ $(p \neq p')$ に対して，類似度の計算をする．類似度の計算には様々なものが存

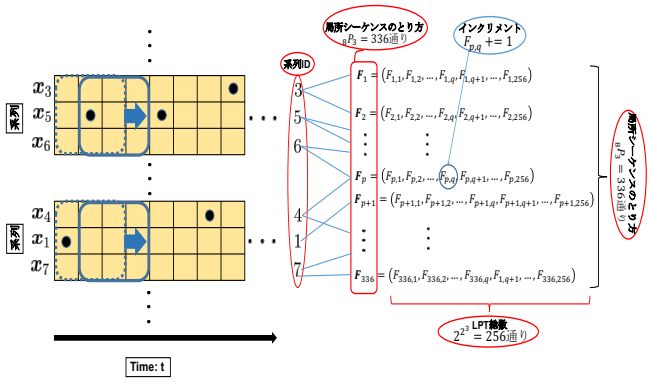

図 4: LPT(Local Pattern Tensol) 度数ベクトルの生成

在するが，一例としてはコサイン類似度を用いたものがあり，それに基づいてクラスタリングを行う．コサイン類似度の計算式は以下の通りである．

$$\cos\left(\boldsymbol{F}_p, \boldsymbol{F}_{p'}\right) = \frac{F_p \cdot F_{p'}}{|F_p|\,|F_{p'}|} = \frac{\sum_{i=1}^{256}\left(F_{p,i} \cdot F_{p',i}\right)}{\sqrt{\sum_{i=1}^{256} F_{p,i}^2}\sqrt{\sum_{i=1}^{256} F_{p',i}^2}}$$

最後に計算した非類似度にもとづいて適切なクラスタリングを行うが，先行研究としては，山川 [1] は LPT ベクトルの各値に対して相互情報量による重みを付けた significance と呼ばれる指標を用いてその値が大きいものから順にクラスタリングすることによって，重要と思われるクラスタを抽出している．適当な基準に則って等価と判断することができる M 個の K-タプルからなる集合を抽出する．最後にまとめとして，多次元時系列データから等価性構造を抽出する一連の過程を Algorithm 1 に簡潔に示す．

Algorithm 1 ES Extraction Method

```
for p = 1 ∼_N P_K do
    F_p を計算
end for
for p = 1 ∼_N P_K do
    for p' = p + 1 ∼_N P_K do
        F_p と F_p' の類似度を計算
    end for
end for
類似度をもとにクラスタリング
```

4 おわりに

本研究では，等価性構造 (ES) が N 次元時系列から非類似度にもとづいて，等価と見做すことの出来る M 要素の K-タプルの集合として取り出す方法を説明するとともに，ES が持つ性質について述べた．

解決が急がれる問題として，K の値が膨大になった場合の組合せ爆発がある．この問題に対して，佐藤らは $K = 2$ から逐次的に ES を抽出し，K の値を増やすアルゴリズムを提案している [6]．これからの展望としては，深層学習の解析や，見まね学習の教師と生徒の次元対応付け等への応用，またその他にも多様なデータセットを対象として等価性構造抽出実験をすることで，一見関係性が無いと思われるもの同士に何らかの関係性を見出すことが考えられる．

謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (ＮＥＤＯ) の委託業務の結果得られたものです．また本研究を遂行するにあたり，ドワンゴ人工知能研究所の皆さまから多大なるご協力をいただきましたことに感謝致します．

参考文献

[1] 山川宏. 局所多次元時系列の関係表現としての性質の実験的検討. *Proceedings of JSAI2013*, No. 3H4-OS-05c-2in, 2013.

[2] Rodolphe Gentili and James Reggia. Imitation learning as cause-effect reasoning. In *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*, Vol. 9782, p. 64. Springer, 2016.

[3] Donald E Knuth, James H Morris, Jr, and Vaughan R Pratt. Fast pattern matching in strings. *SIAM journal on computing*, Vol. 6, No. 2, pp. 323–350, 1977.

[4] JLEKS Lonardi and Pranav Patel. Finding motifs in time series. In *Proc. of the 2nd Workshop on Temporal Data Mining*, pp. 53–68, 2002.

[5] Edgar F Codd. A relational model of data for large shared data banks. *Communications of the ACM*, Vol. 13, No. 6, pp. 377–387, 1970.

[6] 佐藤聖也, 山川宏. 等価性構造保持仮定の下での等価性構造抽出における探索数削減 (to appear). Technical report, 電子情報通信学会コンピュテーション研究会 (COMP), 2016.