高速化に向けた等価性構造の特徴解析

Characteristics Analysis of Equivalence Structures for Speed-Up

佐藤聖也 [1*]　山川宏 [2,3]
Seiya Satoh[1]　Horoshi Yamakawa[2,3]

[1] 産業技術総合研究所
[1] National Institute of Advanced Industrial Science and Technology (AIST)
[2] (株) ドワンゴ ドワンゴ人工知能研究所
[2] Dwango Artificial Intelligence Laboratory
[3] NPO 法人 全脳アーキテクチャ・イニシアティブ
[3] The Whole Brain Architecture Initiative, a specified non-profit organization

Abstract: Time-series data subspaces that can be considered equivalent are called equivalence structures. Extraction of equivalence structures can be applied for various purposes, such as feature extraction, the analysis of neural network behavior, and dimension matching of different datasets. However, the extraction requires comparisons of all possible subspaces of the whole space of a multidimensional time series dataset, and the comparisons cause combinatorial explosion with the number of the dimensions of subspaces. In this paper, we analyze the characteristics of equivalence structures for the development of fast equivalence structure extraction.

1　はじめに

高次元時系列データにおいて等価とみなしうる部分空間の集合を等価性構造と呼ぶ [1]．部分空間同士の局所シーケンスが似ている場合それらの部分空間を等価とみなしうるが，このとき，似ている局所シーケンスの開始時間は部分空間ごとに異なっていてもよい．等価性構造抽出は特徴抽出，ニューラルネットネットワークの挙動解析，異なるデータ間の次元対応付け等，様々な応用が考えられる．異なるデータ間の次元対応付けは，異なるデータの部分空間に共通の特徴を見つけることにより行うため，異なるドメイン間での知識共有を可能とする．例えば，見真似学習は AGI においても重要な課題であるが，現在は教師と生徒の関節等の次元の対応が既知として扱われている [2]．そのため，次元の対応関係が未知である場合や，そもそも体の構造が異なり，次元数も異なることが考えられ，そのような場合，前もって各次元の対応付けを行う必要がある．

等価性構造抽出と類似すると思われる技術の一つとして，データマイニング技術の一種のモチーフディスカバリーがある [3]．モチーフディスカバリーでは，ある次元上に繰り返し現れるパターン（モチーフ）を抽出する．しかし，等価性構造抽出と大きく異なる点は，等価性構造抽出ではある多次元上に現れるパターンと似たパターンが別の多次元上に現れているかに着目する点である．また，モチーフディスカバリーを拡張した方法として，ある次元に現れるパターンと似たパターンが別の次元に現れているかを抽出する方法があるが [4]，この方法は 1 次元のパターンにのみ適用できる方法である．

等価性構造を抽出する方法として，高次元の全体空間から得られる全ての部分空間同士を比較する方法が考えられるが，この場合，部分空間の次元数に伴う，部分空間の比較回数の組み合わせ爆発が問題となる．本稿では，高速な等価性構造抽出アルゴリズム開発に向けた，等価性構造の特徴を解析する．

2　等価性構造（ES）

等価性構造（ES: equivalence structure）は多次元時系列データにおける等価とみなしうる部分空間の集合であり，ES 抽出のためには部分空間を“等価”とみなす尺度が必要である．文献 [1] では，その尺度として，局所シーケンスから得られる分布等価群を用いて等価であるか判定していたが，この方法はバイナリデータにのみ適用できる方法である．連続データにも適用可能な尺度として，本稿では 3.1 節で説明する，局所シー

*連絡先：佐藤聖也，産業技術総合研究所　人工知能研究センター，
〒 135-0064 東京都江東区青海 2-4-7
産総研 臨海副都心センター別館，
E-mail:seiya.satoh@aist.go.jp

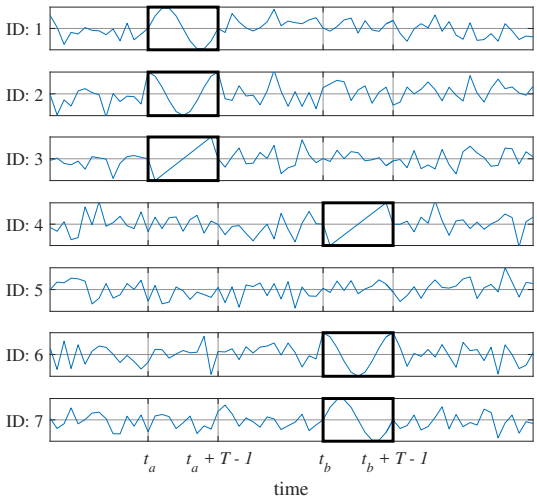

図 1: 3 次元の ES の例（部分空間 $(1,2,3)$ と $(7,6,4)$ が等価性構造である場合）.

ケンスのパターン間の差の平均二乗値を利用した，単純な非類似度を用いる．

図 1 に 3 次元の ES の例を示す．図 1 は，次元 ID1, 2, 3 の部分空間（部分空間 $(1,2,3)$）の時刻 t_a から始まる長さ T の局所シーケンスと，部分空間 $(7,6,4)$ の時刻 t_b から始まる長さ T の局所シーケンスのパターンが酷似している例である．ここでは順序を考慮しているため，部分空間 $(1,2,3)$ の局所シーケンスは部分空間 $(6,7,4)$ や $(4,6,7)$ とは類似していない．また，本稿で用いる非類似度は，3.1 節で説明するように，全ての局所シーケンス同士の比較を基に部分空間同士が等価であるかを判定するため，部分空間 $(1,2,3)$ と $(7,6,4)$ は図 1 に示すようなパターンがあったとしても ES とみなされない可能性があることに留意されたい．

2.1 非類似度の性質

3.1 節で説明する，本稿で用いる非類似度 $d_{\boldsymbol{a},\boldsymbol{b}}$ は以下の 2 種類の対称性がある．ただし，$\boldsymbol{a},\boldsymbol{b}$ は K 個の次元 ID からなる，部分空間を表すベクトルとし，i と j は 1 から K までの任意の整数とする．

$$d_{\boldsymbol{a},\boldsymbol{b}} = d_{\boldsymbol{b},\boldsymbol{a}}, \tag{1}$$

$$d_{(\cdots,a_i,\cdots,a_j,\cdots),(\cdots,b_i,\cdots,b_j,\cdots)} = d_{(\cdots,a_j,\cdots,a_i,\cdots),(\cdots,b_j,\cdots,b_i,\cdots)} \tag{2}$$

ここでは式 (1) の性質を交換対称性，式 (2) の性質を入替対称性と呼ぶこととする．$d_{\boldsymbol{a},\boldsymbol{b}}$ は $\boldsymbol{a},\boldsymbol{b}$ の次元 ID の順序に依存するが，入替対称性は，$\boldsymbol{a},\boldsymbol{b}$ の次元 ID の順序を同じように変更した場合は値が変わらないことを意味している．そのため，既に抽出された ES と入替対称である ES は既出の ES から生成可能である．以下の計算機実験では既出の ES から生成可能な ES は保持しないこととした．

2.2 力任せ探索の比較回数

全ての部分空間を力任せに比較する方法（力任せ探索）では，式 (1) の交換対称性と式 (2) の入替対称性を考慮すると，部分空間の比較回数 n を以下とできる．ただし，$n_p = {}_{K_{\max}}\mathrm{P}_K$，$n_c = {}_{K_{\max}}\mathrm{C}_K = n_p/K!$ とする．

$$n = \frac{n_p(n_p-1)}{2} - \frac{(n_p-n_c)(n_p-n_c-1)}{2}, \tag{3}$$

$$= K! \times n_c^2 - \frac{n_c^2+n_c}{2}, \tag{4}$$

式 (3) の第 2 項は入替対称性を考慮したとき削減できる比較回数を表している．しかし，このように削減したとしても，式 (4) からわかるように，依然として部分空間の次元数に伴う組み合わせ爆発が問題となる．

2.3 等価性構造の性質

この節では以下の仮定を考える．ただし，ある ES の部分空間と等価とみなされる部分空間がその ES の部分空間以外に存在しない場合，その ES を“完全”と呼ぶこととする．また，任意の 2 次元 ES の部分空間の 1 番目と 2 番目の次元 ID をそれぞれの“ペア”と呼ぶこととする．

仮定 1 完全な K 次元の *ES* の全ての部分空間から任意の k 番目の次元 *ID* を取り除いたとき，その $K-1$ 次元の部分空間の集合は，（完全とは限らない）唯一の $K-1$ 次元の *ES* である．

仮定 2 完全な K 次元の *ES* は，ある $K-1$ 次元の *ES* のある部分空間の全ての次元 *ID* とペアである次元 *ID* を加えて生成される集合の中に存在する．

仮定 2 が真であれば，2 次元 ES を用いて 3 次元 ES が得られ，4 次元以上も逐次的に求めることができる．このように ES を求めることで，力任せ探索より大幅に部分空間の比較回数を削減できる可能性がある．以下の計算機実験では仮定 1, 2 が成り立つデータを用いて実験を行った．

3 計算機実験

図 2 に示す，人工的に生成したノイズのない時系列データ $\{x_k^{(t)}, k=1,\cdots,8, t=1,\cdots,36\}$ を用いて実験を行った（$K_{\max}=8, T_{\max}=36$）．局所シーケンスの長さ T は 4 とした．

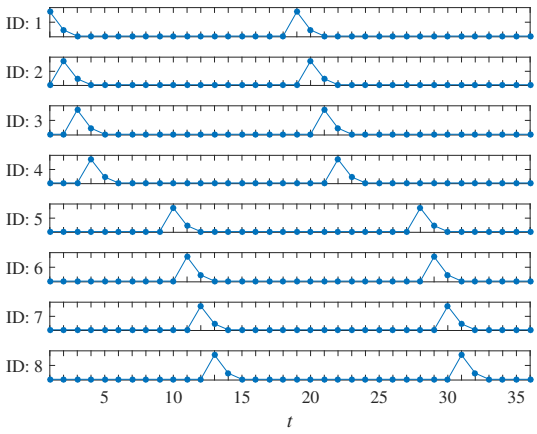

図 2: 人工時系列データ.

3.1 部分空間同士が等価であるかの尺度

部分空間同士が等価であるかの尺度として，ここでは局所シーケンスのパターン同士の差の平均二乗値（MSV: mean-square value）を利用した，単純な非類似度を用いる．具体的には以下のように計算する．ここで，$K_{\max}$ 次元の時系列データ $\{x_k^{(t)}, k = 1, \cdots, K_{\max}, t = 1, \cdots, T_{\max}\}$ が与えられ，以下の，平均を引く正規化を行った局所シーケンス $\boldsymbol{z}_k^{(t)}$ を定義する．ただし，t は時刻，k は次元 ID を示し，$\boldsymbol{c}^{\mathrm{tr}}$ は $\boldsymbol{c}$ の転置，T は局所シーケンスの長さとし，$t = 1, \cdots, T_{\max} - T + 1$ とする．

$$\boldsymbol{z}_k^{(t)} = (z_k^{(t,1)}, \cdots, z_k^{(t,T)})^{\mathrm{tr}} \quad (5)$$

$$= (x_k^{(t)}, \cdots, x_k^{(t+T-1)})^{\mathrm{tr}} - \frac{1}{T}\sum_{\tau=1}^{T} x_k^{(t+\tau-1)} \quad (6)$$

$\boldsymbol{a}$ と $\boldsymbol{b}$ を K 個の時系列データの次元 ID からなるベクトルとしたとき，$\boldsymbol{a}$ の時刻 t_a から始まる局所シーケンスと $\boldsymbol{b}$ の時刻 t_b から始まる局所シーケンス内のパターン間の差の MSV($MSV_{\boldsymbol{a},\boldsymbol{b}}^{(t_a,t_b)}$) は以下である．

$$MSV_{\boldsymbol{a},\boldsymbol{b}}^{(t_a,t_b)} = \frac{1}{TK}\sum_{k=1}^{K}\left|\left|\boldsymbol{z}_{a_k}^{(t_a)} - \boldsymbol{z}_{b_k}^{(t_b)}\right|\right|^2 \quad (7)$$

このとき，部分空間 $\boldsymbol{b}$ の全ての局所シーケンスの内，部分空間 $\boldsymbol{a}$ の時刻 t_a の局所シーケンスと MSV が最も小さいときの MSV （$MSV_{\boldsymbol{a},\boldsymbol{b}}^{(t_a)}$）は以下である．

$$MSV_{\boldsymbol{a},\boldsymbol{b}}^{(t_a)} = \min(MSV_{\boldsymbol{a},\boldsymbol{b}}^{(t_a,t_b)}| \quad t_b = 1, \cdots, T_{\max} - T + 1) \quad (8)$$

次に，$MSV_{\boldsymbol{a},\boldsymbol{b}}^{(t_a)}$ が θ 未満であれば 1, θ を超える場合は 0 を出力する以下の関数を定義する．ただし，$h(c)$ はヘヴィサイドの階段関数とし，$c > 0$ のとき 1, $c < 0$ のとき 0 を出力する．

$$h_{\boldsymbol{a},\boldsymbol{b}}^{(t_a)} = h(\theta - MSV_{\boldsymbol{a},\boldsymbol{b}}^{(t_a)}) \quad (9)$$

上記の関数を用い，部分空間 $\boldsymbol{a}, \boldsymbol{b}$ の非類似度 $d_{\boldsymbol{a},\boldsymbol{b}}$ を以下とする．ただし，$\boldsymbol{1}$ は要素がすべて 1 の $T_{\max} - T + 1$ 行ベクトルとする．

$$d_{\boldsymbol{a},\boldsymbol{b}} = 1 - \frac{\boldsymbol{w_a}^{\mathrm{tr}}\boldsymbol{h}_{\boldsymbol{a},\boldsymbol{b}} + \boldsymbol{w_b}^{\mathrm{tr}}\boldsymbol{h}_{\boldsymbol{b},\boldsymbol{a}}}{\boldsymbol{w_a}^{\mathrm{tr}}\boldsymbol{1} + \boldsymbol{w_b}^{\mathrm{tr}}\boldsymbol{1}}, \quad (10)$$

$$\boldsymbol{h}_{\boldsymbol{a},\boldsymbol{b}} \equiv (h_{\boldsymbol{a},\boldsymbol{b}}^{(1)}, \cdots, h_{\boldsymbol{a},\boldsymbol{b}}^{(T_{\max}-T+1)})^{\mathrm{tr}}, \quad (11)$$

$$\boldsymbol{w_a} \equiv (w_{\boldsymbol{a}}^{(1)}, \cdots, w_{\boldsymbol{a}}^{(T_{\max}-T+1)})^{\mathrm{tr}}, \quad (12)$$

$$w_{\boldsymbol{a}}^{(t_a)} \equiv \frac{1}{T}\sum_{\tau=1}^{T}\sqrt{\sum_{k=1}^{K}\left\{z_{a_k}^{(t_a,\tau)}\right\}^2} \quad (13)$$

ここで，$w_{\boldsymbol{a}}^{(t_a)}$ は局所シーケンス内の各 τ の点 $((z_{a_1}^{(t_a,\tau)}, \cdots, z_{a_K}^{(t_a,\tau)}))$ のノルムの平均であり，$w_{\boldsymbol{a}}^{(t_a)}$ が小さい局所シーケンスよりも大きい局所シーケンスで $h_{\boldsymbol{a},\boldsymbol{b}}^{(t_a)}$ が 1 のときの方が $d_{\boldsymbol{a},\boldsymbol{b}}$ が小さくなる．この非類似度 $d_{\boldsymbol{a},\boldsymbol{b}}$ が十分小さいとき，$\boldsymbol{a}$ と $\boldsymbol{b}$ は“等価”であるとみなし，同じ ES に属すると表現する．このとき，$\boldsymbol{a}$ と $\boldsymbol{b}$ は ES の部分空間と呼び，a_k を部分空間 $\boldsymbol{a}$ の k 番目の次元 ID と呼ぶ．

ES の部分空間は実際には 3 つ以上となることがあり得るため，上記の非類似度を基にクラスタリングして ES を求める．以下の計算機実験では階層クラスタリングの一種の最短距離法を用いた．

表 1: ES 抽出結果.

2 次元 ES 1	{(1, 2), (2, 3), (3, 4), (5, 6), (6, 7), (7, 8)}
2 次元 ES 2	{(1, 3), (2, 4), (5, 7), (6, 8)}
2 次元 ES 3	{(1, 4), (5, 8)}
3 次元 ES 1	{(1, 2, 3), (2, 3, 4), (5, 6, 7), (6, 7, 8)}
3 次元 ES 2	{(1, 2, 4), (5, 6, 8)}
3 次元 ES 3	{(1, 3, 4), (5, 7, 8)}
4 次元 ES 1	{(1, 2, 3, 4), (5, 6, 7, 8)}

3.2 実験結果

力任せ探索により ES を抽出した結果を表 1 に示す．表では，2 次元 ES 1 には部分空間 (1, 2) 等の計 6 個の部分空間が存在することを示している．その他の 2 次元 ES 2, 3, 3 次元 ES 1, 2, 3, 4 次元 ES 1 の部分空間の数はそれぞれ 4, 2, 4, 2, 2 であった．ここでは ES の

表 2: 部分空間比較回数.

ES の次元数	力任せ探索 (A)	仮定 2 を考慮したとき (B)	(A)/(B)
2	1,162	1,162	1
3	17,220	8	2,152.5
4	115,115	1	115,115
5	374,724	0	374,724/0

番号や ES の部分空間の順序に意味はないが，次元 ID の順序には意味があることに留意されたい．

[仮定 1, 2 の検証]

ここではまず，2.3 節の仮定 1 が成り立っているか検証する．力任せ探索により得られた ES の最大次元数は 4 であり，4 次元の ES の数は 1 であった．このとき，4 次元 ES 1 の全ての部分空間から任意の k 番目の次元 ID を取り除いたとき，その 3 次元のグループは唯一の 3 次元 ES に属してることが表 1 からわかる．

同様に各 3 次元 ES の任意の k 番目の次元 ID を取り除くと唯一の 2 次元 ES に属することが表 1 から確認できるため，2.3 節の仮定 1 が成り立っていると言える．また，仮定 2 が成り立つことも確認できる．

[仮定 2 が真のときの部分空間比較回数]

表 2 に仮定 2 が真であることを考慮したときの部分空間比較回数と，力任せ探索の部分空間比較回数を示す．力任せ探索の部分空間比較回数は式 (4) より計算できる．今回の実験では，仮定 2 を考慮して考えられる，3 次元 ES と 4 次元 ES の候補は全て ES となっている．そのため，候補が ES であるか検証するための比較回数は 3 次元 ES 抽出では 8 回（$4 \times 3/2 + 1 + 1$），4 次元 ES 抽出では 1 回となる．一方力任せ探索では 3, 4 次元のときの比較回数はそれぞれ 17,220 回, 115,115 回であった．更に，力任せ探索では 5 次元 ES が存在しないことを確認するため，374,724 回の比較が必要となる．そのため，ES の性質を考慮したときの 3, 4, 5 次元の部分空間の比較回数は，力任せ探索と比べて，それぞれ 2,152.5 倍, 115,115 倍, 374,724/0 倍少なくてすむこととなる．ただし，仮定 2 が真である場合も，2 次元 ES の抽出は力任せ探索と同じ回数（1,162 回）比較する必要がある．

4 むすび

本稿では，高速な等価性構造抽出アルゴリズム開発に向けた，等価性構造（ES）の特徴を解析した．今回の計算機実験で用いた，単純な 8 次元の時系列データでは 2.3 節の仮定 1, 2 を満たすため，仮定 2 を利用して ES を探索すると，ES の次元数が 3 以上では，力任せ探索と比べて 2,152.5 倍以上部分空間同士の比較回数を削減できることがわかった．

今回の計算機実験では仮定 1, 2 が満たされる単純な人工の時系列データを用いたが，今後の課題として，仮定 1, 2 が満たされないデータを含む，多様なデータを用いて実験することや，仮定 1, 2 がどのようなときに満たされるかを解析することが挙げられる．

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ある．

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, Proc. JSAI2013, 3H4-OS-05c-2in, (2013)

[2] Katz, G., Huang, D.W., Gentili, R., Reggia, J.: Imitation Learning as Cause-Effect Reasoning, *International Conference on Artificial General Intelligence*, pp. 64–73 (2016)

[3] Tanaka, Y., Iwamoto, K., and Uehara, K.: Discovery of time-series motif from multi-dimensional data based on MDL principle, *Machine Learning*, Vol. 58, No. 2–3, pp. 269–300 (2005)

[4] Kurasawa, H., Sato, H., Nakamura, M., Matsumura, H.: Online Top-k Similar Time-Lagged Pattern Pair Search in Multiple Time Series, *International Conference on Database and Expert Systems Applications*, pp. 432–441 (2012)