人工知能学会研究会資料
SIG-AGI-006-10

複数のニューラルネットワーク隠れ層出力の等価性抽出の試み

Trial experiments for extracting equivalence structures from the activations of hidden layers in multiple neural networks

高橋 良暢 [1*] 佐藤 聖也 [2] 栗原 聡 [1] 山川 宏 [3,4]
Yoshinobu Takahashi[1] Seiya Satoh[2] Satoshi Kurihara[1] 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: 等価性構造抽出技術は，属性が特定されない多数の系列に対し，その一部の「系列の組」を系列間の関係とみなし，等価と見なしうる系列の組を発見する技術である．たとえば同一もしくは異なる時系列データから，必ずしも同時でない時刻（非同期）に共通した部分系列を含む系列の組を発見する．そこで本稿では，異なりつつも共通部分を含むデータにもとづいて教師あり学習を行った二つの多層パーセプトロン（MLP）について，その隠れ層に含まれる共通部分を抽出するための準備をすすめた．等価性構造を抽出しうる二つのデータ内に共通する振る舞いが含まれることを確認し，その上で等価性構造抽出技術を適用すれば，等価性構造が得られうることを示す．

1 はじめに

等価性構造は属性が特定されない多数の系列に対し，その一部の「系列の組」を系列間の関係とみなし，等価と見なしうる系列の組を発見する技術である [1]．ここでは複数の「系列の組」同士の間において十分に類似した部分系列 (共通成分) が含まれていればそれらを等価と見做す．本技術の応用分野としては特徴抽出や，現在既知として扱われている見まね学習の前処理としての教師と生徒の次元の対応付け [2] などが期待される．なお類似する問題として，単一系列から特定のパターンを抽出するモチーフディスカバリーがあるが [3]，等価性構造抽出では，ある多次元上に現れるパターンが別の多次元上で現れるかを確認する点で異なる．

近年，ニューラルネットワークの応用が様々な分野において進展しているため，その隠れ層の表現を人間が理解することは重要な技術テーマとなってきている．これを解決するアプローチとして，異なるデータから学習したニューラルネットワークの隠れ層の表現における潜在的な共通成分 (つまり等価性構造) を抽出することは有効なアプローチであると考えられる．

そこで最近我々は，二つの多層パーセプトロン（MLP）にそれぞれ異なる動画を構成する静止画像を学習させ，その後，二つの MLP 間でその隠れ層の表現を分析するために，等価性構造抽出を用いる試みを開始したのでその結果を報告する．

本稿では，まず 2 節で等価性構造の定義を簡潔に述べる．詳しい定義に関しては，参考文献 [4] や [5] を参照されたい．続く 3 節では，本稿で行った計算機実験の内容を説明するとともに，等価性構造抽出の対象となるニューラルネットワーク，またその隠れ層について述べる．4 節では，等価性構造抽出を行った結果について記すとともに，その結果について考察する．5 節では，4 節までに述べた内容をまとめるとともに，現状の課題と今後の展望について述べる．

2 等価性構造抽出

2.1 概要

等価性構造抽出技術は与えられた N 個の系列から，等価性構造を発見する手法である．等価性構造は，等価と見なしうる同じ系列数 K で構成される「系列の組」からなる’集合’である．さらに系列の組は，長さ K のタプルが要素となるような集合である．長さ K のタプルを以後 K タプルと記す．K タプルは図 1 の出力にあ

*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@ics.lab.uec.ac.jp

るような，N 個の各系列に対して割り振られた ID を要素とし，順序を考慮するものである．等価性構造は入力である N 個の系列からいくつ取得できるかは不明であり，等価性構造の要素である K タプルは，異なる K タプルの部分系列における非類似度を基に等価とみなすことできるものが要素となる [4]．このとき比較するそれぞれの K タプルの部分系列の開始点は異なっていてよい．図 1 の例では，入力として 7 本以上の系列が与えられ，これらには $\#1, \#2, \ldots, \#7, \ldots$ と ID が割り振られている．これに対して，図の出力では，三つ以上の等価性構造が抽出され，その中の一つとしてタプル $< \#1, \#2, \#3 >$ と，タプル $< \#7, \#6, \#4 >$ が等価と見なされている．タプルのサイズや構成する ID が同一でも，その順番が異なれば同一の等価性構造に属するとは限らない．

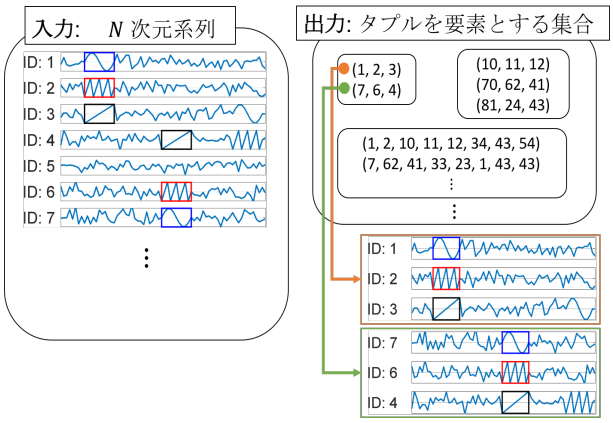

図 1: 等価性構造抽出の定義

2.2 非類似度関数による等価性の判定

タプル間の部分系列同士が等価であるかどうかの基準として，ここでは部分系列同士のパターンにおけるユークリッド距離の平均二乗値 (Mean-Square Values: MSV) を使用した，単純な非類似度を用いた．$[N]$ は $\{1, 2, \ldots N\}$ を表すものとし，与えられた N 次元の系列を $\left\{x_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(T)}\right)\right\}_{i \in [N]}$ とおく．また，ID$k \in [K]$ の部分系列を

$$\begin{aligned} \boldsymbol{z}_k^{(t)} &= \left(z_k^{(t,1)}, z_k^{(t,2)}, \ldots, z_k^{(t,\tau)}\right)^{tr} \\ &= \left(x_k^{(t)}, \ldots, x_k^{(t+\tau-1)}\right)^{tr} - \frac{1}{\tau}\sum_{t'=1}^{\tau} x_k^{(t+t'-1)} \end{aligned} \quad (1)$$

とする．このとき，$t\,(= 1, 2, \ldots, T-\tau+1)$ は時間，τ は部分系列の長さである．また，$\boldsymbol{u}_1$，$\boldsymbol{u}_2$ を K タプル，$u_{1,k}$，$u_{2,k}$ を $\boldsymbol{u}_1$，$\boldsymbol{u}_2$ の k 番目の要素とする．ここから，次のように $\boldsymbol{u}_1$，$\boldsymbol{u}_2$ における MSV を考える．

$$MSV_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1,t_2)} = \frac{1}{\tau K}\sum_{k=1}^{K}\left\|z_{u_1,k}^{(t_1)} - z_{u_2,k}^{(t_2)}\right\|^2 \quad (2)$$

また，$MSV_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1,1)}, \ldots, MSV_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1,T-\tau+1)}$ の最小値を $MSV_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1)}$ とし，以下の様に定義する．

$$MSV_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1)} = \min\left(MSV_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1,t_2)} | t_2 = 1, \ldots, T-\tau+1\right) \quad (3)$$

次に，以下で表される二値関数を導入する．

$$h_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1)} = h\left(\theta_{MSV} - MSV_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t_1)}\right) \quad (4)$$

式 3 は $MSV_{u_1,u_2}^{(t_1)} > \theta_{MSV}$ となるときに 0 を出力し，$MSV_{u_1,u_2}^{(t_1)} < \theta_{MSV}$ となるときに 1 を出力するようなヘビサイドの階段関数である．以上から，二つの K タプル $\boldsymbol{u}_1$，$\boldsymbol{u}_2$ に対して，

$$d_{\boldsymbol{u}_1,\boldsymbol{u}_2} = 1 - \frac{1}{\beta}\sum_{t=1}^{T-\tau+1} w_{\boldsymbol{u}_1}^{(t)} h_{\boldsymbol{u}_1,\boldsymbol{u}_2}^{(t)} + w_{\boldsymbol{u}_2}^{(t)} h_{\boldsymbol{u}_2,\boldsymbol{u}_1}^{(t)} \quad (5)$$

で表される非類似度を計算する．ここで，

$$w_{\boldsymbol{u}_1}^{(t_1)} = \frac{1}{\tau}\sum_{t'=1}^{\tau}\sqrt{\sum_{k=1}^{K}\left\{z_{u_k^{(1)}}^{(t_1,t')}\right\}^2} \quad (6)$$

であり，$\beta = \sum_{t=1}^{T-\tau+1} w_{\boldsymbol{u}_1}^{(t)} + w_{\boldsymbol{u}_2}^{(t)}$ である．$w_{\boldsymbol{u}_1}^{(t)}$ は部分系列に重みを加えることを目的としており，この値が大きければ大きいほど，$MSV_{u_1,u_2}^{(t_1)} < \theta_{MSV}$ のときにタプル同士を等しいと判定する．

以上の基準をもって，MATLAB R2016b Statistics and Machine Learning Toolbox version を使用して，階層的クラスタリングを行った．この際の閾値を θ_{th} で表す．

3 計算機実験による検討

今回はまず，ニューラルネットワークの内部状態の解析に等価性構造技術を利用できる可能性を見出すために，データセット (DS) として異なる動画の各フレームを構成する静止画像を教師あり学習させた二つの多層パーセプトロン（MLP）を準備した．以降これらの MLP をモデル 1 およびモデル 2 と呼ぶ．

検討内容としては，二つのモデルの隠れユニットの出力ベクトルの時系列についての分析を行う．まず，二つのモデルの隠れ層に共通に含まれる振る舞いを線形変換によって抽出できることを確認し，その上で等価性構造抽出技術を適用すれば等価性構造を特定しうるかを検討した．

3.1　データセットに用いた動画像

今回は Disentanglement testing sprites dataset[6] を用いた．離散的な動きを含むサイズ 64 × 64 ピクセルの画像 737,280 枚から，図 2 に一端を示すような画像 752 枚で構成される二つの連続的な動画像を構成し，データセットとした．具体的には，一つ目の動画像（DS1）は白色の正方形が 64 × 64 ピクセルの画像内を上下に蛇行しつつ左方から右方へ推移する軌跡を描き，右端に辿り着くと始点に戻ってくる．その後始点で大きさの収縮を経て，同様の蛇行を繰り返して始点に戻ってくるものである．二つ目（DS2）は描く軌跡は DS1 と同一であるが，6 フレームごとにハート，楕円，正方形と形を変えながら推移するものである．DS2 は形が変化する分 DS1 より複雑である．この二つの動画像を MLP の入力とする．また教師信号として，画像における図形の色，形，大きさ，角度，X 座標，Y 座標を表す六つ組のベクトルが与えられる．

3.2　ニューラルネットワークモデル (MLP)

実験には，入力層 (64×64 ユニット)-隠れ層（N_m ユニット）-出力層（6 ユニット）の 3 層で構成される，二つの多層ニューラルネットワーク (MLP) を用い，各モデルをモデル 1，モデル 2 とした．

隠れユニット数 N_m（ここで $m \in \{1, 2\}$ はモデル ID) は出力層における教師信号に対する平均二乗誤差 (MSE) を十分に小さくする値を利用することにした．隠れユニット数を 1 から順に増やし，各ユニット数において 10 回 MSE の計測を行い，MSE が 0.05 以下となった際の隠れユニット数として，それぞれ $N_1 = 2$, $N_2 = 4$ を採用した．

4　計算機実験

4.1　線形変換による前処理

隠れ状態の表現には線形変換に対する任意性があり，そのままでは二つのモデル間でユニットを対応づけできない． そこで今回は，モデル 2 の隠れ層出力をモデル 1 の隠れ層出力に対応させる線形変換を行った．

学習後のモデル 1 の隠れユニットから得られる系列は，DS1 をモデル 1 に入力することにより得た．これらの系列を長さ $T = 752$ の $\boldsymbol{x}_1$,$\boldsymbol{x}_2$ とする．学習後のモデル 2 の隠れユニットから得られる系列は，DS2 をモデル 2 に入力することにより得た．その後これらの系列を線形変換し，変換後の四つの系列を長さ $T = 752$ の $\boldsymbol{x}_3$,$\boldsymbol{x}_4$,$\boldsymbol{x}_5$,$\boldsymbol{x}_6$ とした．$\boldsymbol{x}_1$ から $\boldsymbol{x}_6$ のうち，$t = 300, 301, \ldots, 400$ までを図 3 に示す．

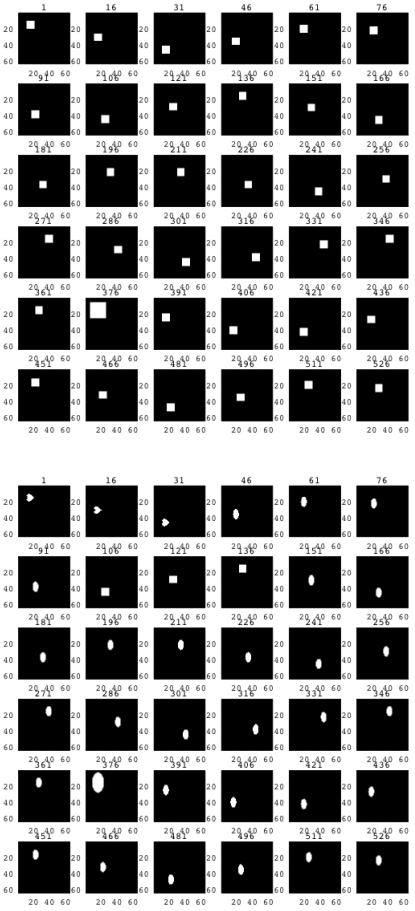

図 2: データセット (DS) の一部： 上図:モデル 1 に用いた DS1，下:モデル 2 に用いた DS2．二つの DS で軌跡とサイズ変化は共通しているが，DS2 では形が変化している．

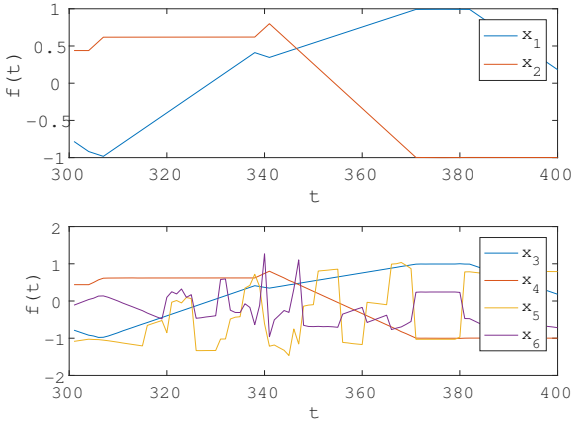

図 3: 上：モデル 1 の隠れ層 (2 ユニット) 出力，下：モデル 2 の線形変換済みの隠れ層 (4 ユニット) 出力．時刻 300〜400 のみ表示

全ての系列 $\boldsymbol{x}_1$ から $\boldsymbol{x}_6$ はそれぞれ ID#1 から #6 を割り当てた．また，各系列は平均が 0，分散が 1 になるように正規化した．

こうして，異なるデータを学習した二つの多層パーセプトロン（MLP）の隠れ層において，二つのデータ内の共通する振る舞いを線形変換によって抽出する．

4.2 実験： 等価性構造の抽出確認

次に，適切に線形変換を施した上であれば，等価性構造抽出技術を適用することで適切な等価性構造が得られるか否かを確認する．本実験において等価性構造抽出用いたパラメータは，$\tau = 50$,$\theta_{MSV} = 0.05$，$\theta_{th} = 0.3$ である．

その結果，モデル 1 における 2 タプル $\langle \#2, \#1 \rangle$ と，モデル 2 における 2 タプル $\langle \#4, \#3 \rangle$ を要素とする集合が等価性構造として抽出された．この結果を図 4 に示す．

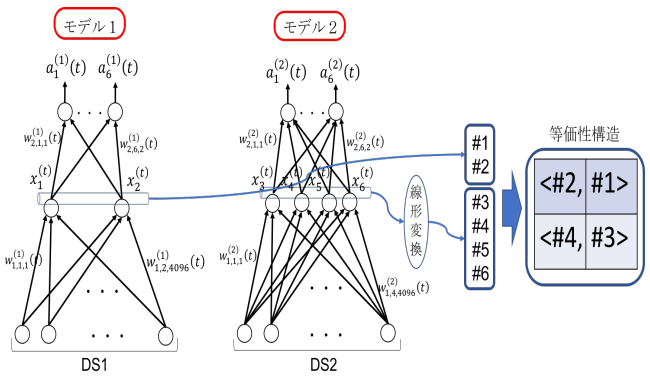

図 4: 実験設定と結果：モデル毎に入力は動画像 DS1,DS2、出力 $a_1^{(m)}(t), \ldots, a_6^{(m)}(t)$ は画像中の図形の (色， 形，大きさ，角度，X 座標，Y 座標)

以上より，二つのデータ内に含まれる共通する振る舞いを等価性構造として抽出できることを確認した．

4.3 考察

モデル 1 は図形の X 座標，Y 座標の変化と，図形の大きさの変化の三つの属性の変化を学習した．モデル 2 はモデル 1 が学習する変化に加えて，一定時刻ごとに変化する図形の形を学習した．二つのモデルの隠れユニット群の間で，タプル $\langle \#4, \#6 \rangle, \langle \#5, \#6 \rangle, \langle \#6, \#4 \rangle, \langle \#6, \#5 \rangle$ がモデル 1 の系列を要素とするタプルと等価とみなされなかった．このことから，系列 #5，系列 #6 は図形の形を認識する役割を持っていると考えられる．つまり一方の MLP に与えたデータのみに含まれる変化は得られた等価性構造以外の系列に対応している．このことから等価性構造は共通する変化を捉えていると考えられる．

5 まとめ

本稿では等価性構造抽出技術をニューラルネットワークの内部表現分析に適用する準備として，異なるデータを学習した二つの多層パーセプトロン（MLP）の隠れ層において，二つのデータ内の共通する振る舞いを，線形変換によって抽出できることを確認した．その上で等価性構造抽出技術を適用すれば等価性構造が得られうることを示した．これにより，ニューラルネットワークの内部状態の解析に等価性構造技術を利用できる可能性が見出された．

今回は，二つのデータセット中の共通要素が同期しているために，等価性構造抽出技術の強みを発揮できる課題となっていないが，今後は非同期なデータセットに拡張したい．また特定の等価性構造に含まれる「系列の組」が二つ以上になった場合には，その系列組数の自乗に比例した数の線形変換を考慮する必要があるので，その増大への対応についても検討したい．

謝辞

本研究にあたり，ドワンゴ人工知能研究所および電気通信大学栗原研究室の皆様からご協力をいただきましたことに感謝致します．また本研究の一部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) の委託業務として行われました．

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. In *Proc. of JSAI2013*, No. 3H4-OS-05c-2in. JSAI, 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] S. Lonardi J. Lin, E. Keogh and T. Pranav. Finding motifs in time series. In *Proc. of the 2nd Workshop on Temporal Data Mining*, pp. 53–68, 2002.

[4] Seiya Satoh and Hiroshi Yamakawa. Incremental extraction of high-dimensional equivalence structures. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pp. 1518–1524, 2017.

[5] 高橋良暢, 佐藤聖也, 山川宏. 等価性構造抽出技術の定式化. 第四回汎用人工知能研究会, No. SIG-AGI-004-03. JSAI, 2016.

[6] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.