Vol.2018-ICS-190 No.3
2018/3/2

複数のニューラルネットワークの隠れ層出力に対する等価性構造抽出

高橋 良暢[1,a)]　佐藤 聖也[2,b)]　栗原 聡[1,c)]　山川 宏[3,4,d)]

概要：近年，計算機の計算能力が飛躍的に向上したことに加えて，Deep Learning の有用性が広く認知されてきたことにより，ニューラルネットワークの産業応用が医療，製造業，エンターテインメントなど様々な分野で大きく進展している．しかし，ニューラルネットワークが出力を導き出すまでの過程がブラックボックスとなっていることから，その内部表現の理解，即ち隠れ層の表現を人間が理解することは重要である．そこで我々は異なるデータセットで学習を行った二つのニューラルネットワークの隠れ層出力に対して，時系列データマイニングの一つである等価性構造抽出技術を用いた実験を行った．等価性構造抽出技術は，属性が特定されない多数の系列が与えられたときに，それらの系列の ID で構成される組（タプル）を系列間の関係とみなし，その系列の組同士について等価な関係を発見する技術である．本論文では，二つの多層パーセプトロン (MLP) にそれぞれ異なりつつも，差分が明確な動画像を学習させ，その後，二つの MLP 間でその隠れユニットの出力ベクトルの時系列を比較して，等価性構造を抽出する実験を，データセットの種類に応じて二通りの実験を行った．これにより，MLP の隠れユニットが特定の属性に反応しているかどうかを確認し，その上で適切な前処理を行えば，等価性構造が取得できることを示した．また，各 MLP が個別に学習するデータセットの関係によって，等価性構造の抽出がどのような意味を持つかを検討する．

Extracting Equivalence Structures from the Activations of Hidden Layers in Multiple Neural Networks

Yoshinobu Takahashi[1,a)]　Seiya Satoh[2,b)]　Satoshi Kurihara[1,c)]　Hiroshi Yamakawa[3,4,d)]

1. はじめに

近年，計算機の計算能力が飛躍的に向上したことに加えて，Deep Learning[1] の有用性が広く認知されてきたことにより，ニューラルネットワークの産業応用が医療，製造業，エンターテインメントなど様々な分野で大きく進展している [2][3][4]．しかし Deep Learning を超えるような発明のためには，現在ブラックボックスとなっているニューラルネットワーク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いる．本論文では，時系列データマイニングにおいて新たな知見の発見が期待される，等価性構造抽出技術を用いることで，共通する部分を持ちながらも異なる，二つのデータセットを学習したニューラルネットワークの，隠れ層の出力に共通する振る舞いを抽出した．

等価性構造抽出技術は，多次元の系列データの間から，等価な関係を見出す技術である [5]．本技術は，多次元系列上と，別の多次元系列上に現れる時系列上のパターンに着目し，共通のパターンをもっている系列の ID で構成されるタプルを見つけるものである．等価性構造抽出技術の入力は，系列の集合であり，出力は等価性構造と呼ばれる集

1 電気通信大学
University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
2 産業技術総合研究所　臨海副都心センター
National Institute of Advanced Industrial Science and Technology, Koto, Tokyo 135–0064, Japan
3 （株）ドワンゴ　ドワンゴ人工知能研究所
DWANGO Co., Ltd. Bunkyo, Tokyo 113–0033, Japan
4 NPO 法人　全脳アーキテクチャイニシアティブ The Whole Brain Architecture Initiative, a specified non-profit organization, Japan
a) ytakahashi@ics.lab.uec.ac.jp
b) seiya.satoh@aist.go.jp
c) kuri@acm.org
d) hiroshiyamakawa@dwango.co.jp

合である．等価性構造は K 個の系列 ID で構成される，K-タプルを要素とする集合であり，どの K 系列で構成されるタプルを等価と見做すことができるのかを示す．また，どのタプルとどのタプルが等価か決定する方法としては，K 次元系列の部分系列にもとづいて決定する．

ニューラルネットワークの隠れ層についての研究には，立石と山崎による，手書き文字認識におけるニューラルネットワークの入力層から隠れ層への重み分布による隠れ層の考察 [6] や，D. L. K Yamins らによる，convolutional neural network を実際の猿の高次視覚野と比較することで，両者のを比較することで理解を得ようとしてる研究がある [7]．また，DeepLearning の隠れ層に関して，Kavukcuoglu ら [8] は，画像認識で高い精度を出した Deep Learning の隠れ層が抽出している特徴量を分析し，入力層に近い隠れ層側では，入力データの線分や点などの抽象的な特徴が捉えられていることを示した．また，Zeiler らは，出力層に近い隠れ層が学習した特徴量は，より具体的な特徴であるということを明らかにしている [9]．

しかし，これらの研究は全てユニット間の重み係数に着目した研究であり，隠れユニットの役割について考える際に，隠れユニットの出力に着目している研究はほとんど存在していない．

また等価性構造抽出に類似した研究としては，モチーフディスカバリーが挙げられる．モチーフディスカバリーは，ある単一の系列上に現れる特定のパターンが，同一の系列上に現れるかどうかを探索する手法であり，広く研究されている．しかし，モチーフディスカバリーはある系列上で現れたパターンを他の系列上で探索することができない点で等価性構造抽出と異なる．また，モチーフディスカバリーは多次元上に現れるパターンを同時に捉えることができず，系列間の関係に着目している等価性構造抽出とはその点でも異なるといえる．[10][11][12][13][14] また，モチーフディスカバリーは分類，クラスタリング，可視化，ルール発見などの高度な解析において広く応用されている [15][16][17][18]．

本研究では，等価性構造抽出技術 [5] を用いて，異なるデータセットの属性を教師あり学習した二つのニューラルネットワークから，特定の属性に反応するような隠れユニット群を見つけ出すことを目的とする．具体的には，異なりつつも共通部分が判明しているデータセットを用いて学習を行った二つのニューラルネットワークの隠れ層に対して，データセットの共通部分の変化に反応するような隠れユニット群を抽出し，隠れ層の役割を理解することである．

本稿の構成について述べる．以下，第二章では等価性構造抽出技術の問題設定を説明する．第三章では本稿で行った計算機実験について述べる．続く第四章では実験の結果と考察を述べ，最後の第五章で結論と今後の課題について述べる．

2. 等価性構造抽出技術

等価性構造は，与えられた $N \in \boldsymbol{N}^+$ 個の系列から抽出することのできる，長さ $K < N$ の等価と見做し得る複数のタプルを要素とする集合である．また，これを抽出する技術を等価性構造抽出技術と呼ぶ．長さ K のタプルは K 次元の系列を表し，これを K-タプルと呼ぶ．複数の K-タプルで構成される等価性構造を K 次元 ES，もしくは K-ES と呼ぶ．等価性構造抽出の概念を図に表したものを下図 1 に示す．各系列には ID を割り当て，各 ID を#でによって番号付けしている．図 1 では，$\#1, \ldots, \#8$ までの $N = 8$ 次元の系列が与えられ，青，赤，黒等の色付きで表されている部分系列の比較をしている．このとき，赤と赤，黒と黒のように単一の部分系列同士を比較しているのではなく，複数の部分系列をひとまとめとして扱い，それらを同時に比較しているところに等価性構造抽出の特徴があるといえるが，詳しい説明は後の節に譲る．出力として，2 次元の ES が三つ $(\{\langle \#1, \#2\rangle, \langle \#8, \#7\rangle\}, \{\langle \#1, \#3\rangle, \langle \#8, \#5\rangle, \langle \#8, \#6\rangle\}, \{\langle \#2, \#3\rangle, \langle \#7, \#5\rangle, \langle \#7, \#6\rangle\})$，3 次元の ES が一つ $(\{\langle \#1, \#2, \#3\rangle, \langle \#8, \#7, \#5\rangle, \langle \#8, \#7, \#6\rangle\})$ 抽出されている．

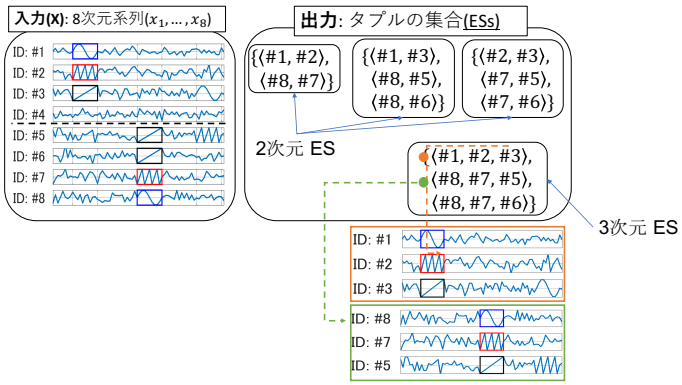

図 1　等価性構造抽出の概念図

N 次元系列を考える．系列自身を $\boldsymbol{x}_i \in X = \{\boldsymbol{x}_i | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ とおく．また，系列における時間を $t = 1, 2, \ldots, T$ とおく．このとき，時刻 t における x_i の値を $x_{t,i}$ と表すことで，系列 ID$\#i$ である系列の，ある時刻における値がただひとつに決まる．

等価性構造について詳しく説明していく．等価性構造抽出技術はある基準を以て，等価と見做すことができる複数のタプルを要素とする集合である．この結果抽出されたものが等価性構造である．

N 次元系列の各系列 $\boldsymbol{x}_i$ に割り当てられる系列 ID：$\#i$ を値としてもつ関係データについて考える．関係データの組（行）として，系列 ID を要素として持つタプルを定義

し，これを行タプルと呼ぶ．行タプルの下図は M とし，その指定子を $m\,(1 \leq m \leq M)$，行タプルを $\boldsymbol{v}_m$ で表す．

次に等価性構造の列について考える．列数を K であり，各列の要素が M 個全てのタプルの $k\,(1 \leq k \leq K)$ 番目の系列 ID で構成される M-タプルが等価性構造の各列となる．この M-タプルを射タプル（morphism tuple）と呼び，$\boldsymbol{v}^{(k)}$ で表す．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持つ系列 ID で構成される．以上から，等価性構造を構成する特定の系列 ID は $v_m^{(k)}$ で表される．以上のような関係データを等価性構造と呼び，$\boldsymbol{S}^{(K)} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M\}$ で表される．

2.1 部分系列が等価であると見做す定義及び非類似度関数

等価性構造抽出の難しさの一つに，複数次元の系列の比較が，単に一次元の系列同士の比較を繰り返しすだけでは実現できないという点がある．これはあるタプル同士が等価だと判定できたとしても，そのタプルを基準としたそれより高次元のタプルが等価あることは保証されないことによる．

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-Squared-Value）を使用する単純な（非）類似度を用いた．以降，$[N]$ は $\{1, 2, \ldots N\}$ を表すものとし，与えられた N 次元の系列を $\left\{\boldsymbol{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(T)}\right)\right\}_{i \in [N]}$ とおく．また，ID$k \in [K]$ の部分系列を

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

で表す．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．$\boldsymbol{c}^{tr}$ はベクトル $\boldsymbol{c}$ の転置である．

次に，二つの異なる K 次元の部分系列間の MSV について考える．二つの異なる K-タプル $\boldsymbol{v}_1$，$\boldsymbol{v}_2$ で表される K 次元系列間の MSV は以下の式 2 で表される．

$$MSV_{v_1,v_2}^{(t_1,t_2)} = \frac{1}{\tau K}\sum_{k=1}^{K}\left|z_{v_1,k}^{(t_1)} - z_{v_2,k}^{(t_2)}\right|^2 \tag{2}$$

このとき，$v_{1,k}$，$v_{2,k}$ はそれぞれ $\boldsymbol{v}_1$，$\boldsymbol{v}_2$ の k 番目の要素である．また，$MSV_{\boldsymbol{v}_1,\boldsymbol{v}_2}^{(t_1,1)}, \ldots, MSV_{\boldsymbol{v}_1,\boldsymbol{v}_2}^{(t_1,T-\tau+1)}$ の最小値を $MSV_{\boldsymbol{v}_1,\boldsymbol{v}_2}^{(t_1)}$ とし，以下の様に定義する．

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

次に，以下で表される二値関数を導入する．

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

式 4 の関数 h は $MSV_{v_1,v_2}^{(t_1)} > \theta_{MSV}$ となるときに 0 を出力し，$MSV_{v_1,v_2}^{(t_1)} < \theta_{MSV}$ 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 $\boldsymbol{v}_1$，$\boldsymbol{v}_2$ に対して，

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

で表される非類似度を計算する．ここで，$w_{\boldsymbol{v}_1}^{(t)}$ と β はそれぞれ

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

$$\beta = \sum_{t=1}^{T-\tau+1} w_{\boldsymbol{v}_1}^{(t)} + w_{\boldsymbol{v}_2}^{(t)} \tag{7}$$

である．$w_{\boldsymbol{v}_1}^{(t)}$ 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，$MSV_{v_1,v_2}^{(t_1)} < \theta_{MSV}$ のときにタプル同士を等しいと判定する．また，β は非類似度関数の値を $[0, 1]$ に制限するために用いる．

式 5 の値を基に，MATLAB R2017b Statistics and Machine Learning Toolbox version11.1 を使用して，最短距離法による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．

2.2 探索アルゴリズム

等価性構造の探索法には，現在，二種類の探索法が提案されている．1 つめは，力任せ探索（Brute-Force Search）である．Brute-Force Search は K-タプルの要素である K 個の系列の全ての組み合わせについて比較を行うことで K 次元 ES を得る手法である．二つ目は佐藤らによる，ESIS (Equivalence Structure Incremental Search)[19] を用いた．ESIS は等価性構造の全探索（Brute-Force Search）が K が大きくなるにしたがって計算量が爆発し，現実的な時間では不可能であることから考案された，等価性構造抽出を行うための探索法である．

3. 計算機実験

本研究における実験について説明する．本実験では，データセットとして二つの異なる動画における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）を実験 1，実験 2 においてそれぞれ二つずつ準備した．二つの実験の違いは，実験 1 では，2 つのデータセットにおける変化の差分が，包含関係になっているのに対して，実験 2 では，2 つのデータセットにおける

属性の変化が，一部を共有している関係であるということである．本実験では，これらのデータセットのを学習したMLP における，隠れユニットの出力ベクトルの時系列について分析を行い，どのような操作を前処理として加えれば等価性構造が抽出できるのかを検討する．その上で，通常属性が分散して認識される MLP の隠れユニットに対して，等価性構造が適用できるかを検討していく．

3.1 実験 1

3.1.1 データセットの詳細

本稿で使用するデータセットは，DeepMind 社による，Disentanglement testing sprites dataset[20] を用いた．本データセットは，離散的な動きを含むサイズ 64×64 ピクセルの画像 737,280 枚で構成される．64×64 ピクセルの連続的な動きをする 752 枚の画像を用いて新たに動画像を構成した．

それぞれのフレームを構成する画像の各ピクセルを入力データとし，その際，画像が持つ色，形，大きさ，向き，X 座標，Y 座標の六つ組 $\langle y_1, y_2, y_3, y_4, y_5, y_6 \rangle$ を教師あり学習する．また，各 $y_i\,(i = 1, 2, \ldots, 6)$ は各々 $y_1 \in \{white\}$，$y_2 \in \{square, eclipse, heart\}$，$y_3 = [0.5, 0.6, \ldots, 1.0]$，$y_4 = [0, 2\pi]$，$y_5 = [0, 1]$，$y_6 = [0, 1]$ をもつ．以下では，各実験における詳細を説明する．

一つ目の動画像 (データセット 1) は，64×64 ピクセルの画像内を白色の正方形が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拡大・縮小を経て，同様の蛇行を繰り返して始点に戻ってくるものである．二つ目の動画像 (データセット 2) は，描く軌跡は データセット 1 と同様であるが，6 フレームごとにハート，楕円，正方形と画像中の図形が形を変えながら推移するものとなっている．つまり，データセット 1 は y_3, y_5, y_6 の値が変化し，データセット 2 は y_2, y_3, y_5, y_6 の値が変化する（表 1 参照）．データセット 2 は y_2（形）が変化する分データセット 1 より複雑である．

表 1 データセット 1 とデータセット 2 の属性変化

属性＼データセット	データセット 1	データセット 2
y_1 (Color)		
y_2 (Shape)		変化
y_3 (Scale)	変化	変化
y_4 (Orientation)		
y_5 (Position X)	変化	変化
y_6 (Position Y)	変化	変化

3.1.2 モデルとして用いたニューラルネットワーク

実験 1 では，入力層（64 × 64 ユニット）-隠れ層（N_m ユニット）-出力層（最大 6 ユニット）の 3 層で構成される，二つの MLP を用い，データセット 1 で学習を行ったものをモデル 1，データセット 2 で学習を行ったものをモデル 2 とした．また隠れユニット数 N_m（ここで，$m \in \{1, 2\}$ はモデル ID）の決定方法については，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数を利用することとした．具体的には，隠れユニット数を 1 から順に増やして学習するが，各ユニット数において 1 0 回計測を行い，MSE が 0.05 以下になった際の隠れユニット数を隠れユニット数として採用し，各モデルそれぞれ $N_1 = 2$，$N_2 = 4$ を採用した．

3.1.3 線形変換による前処理

隠れユニットの表現は属性が分散しており，どの隠れユニットが形や角度に反応しているかということは明確に定まっていない．そのため，等価性構造抽出技術を用いて等価な隠れユニット群を発見するためには，適切なある行列によって，系列の変換を行わなければならない．そこで，実験 1 では，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，データセット 1 をモデル 1 に入力することにより得た．これらの系列の長さ $T = 752$ 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，データセット 2 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 $\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6$ とする．

3.1.4 結果

上述のように適切な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等価性構造が抽出得られるか否かを確認した．本実験において，等価性構造抽出に用いたパラメータは，$\tau = 50, \theta_{MSV} = 0.01, \theta_{th} = 0.1$ 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．

表 2 抽出された等価性構造（実験 1 ）

$\boldsymbol{S}_1^{*(2)}$	$v^{(k)}$
	$\langle \#2, \#1 \rangle$
	$\langle \#4, \#3 \rangle$

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．

3.1.5 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 が学習する三つの変化に加えて，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，2-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 ID#5，#6 の系列は図形の形を認識する役割を強く持っていることとと考えられる．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．この結果から，もしニューラルネットワークの学習において，上手く学習ができなかった場合に，学習したい属性を細分化したものを学習できているモデルがあるならば，そのモデルと学習が上手くいかないモデルとの間で等価性構造を抽出することで，どの属性の学習が上手くいっていないのかがわかる可能性があることがわかった．

3.2 実験 2

実験 1 におけるデータセット 1 とデータセット 2 は，データセットの属性の変化が包括関係にあった．実験 2 では，データセットごとに変化する属性は把握できるが，片方のデータセットがもう片方データセットと包含関係にはなっていないため，隠れユニット同士の対応関係を確定することができない状況で，どのような操作をすれば等価性構造が抽出できるかを検討する．

3.2.1 データセットの詳細

1 つめの動画像は, 実験 1 におけるデータセット 2 と同様のものを使用した．二つ目の動画像（データセット 3）は，図形が描く軌跡や形はデータセット 1 と同様だが，それに加えて図形が一定の周期で回転する変化（y_4 の変化）があるものを作成し，使用した．即ち，データセット 2 は y_2, y_3, y_5, y_6 が変化し，データセット 3 は y_3, y_4, y_5, y_6 が変化する（表 3 参照）．

表 3 データセット 2 とデータセット 3 の属性変化

属性＼データセット	データセット 2	データセット 3
y_1 (Color)		
y_2 (Shape)	変化	
y_3 (Scale)	変化	変化
y_4 (Orientation)		変化
y_5 (Position X)	変化	変化
y_6 (Position Y)	変化	変化

3.2.2 モデルとして用いたニューラルネットワーク

実験 2 においても，実験 1 と同様に入力層（64×64 ユニット）-隠れ層（N_m ユニット）-出力層（最大 6 ユニット）の 3 層で構成される，二つの MLP を用い，各モデルをモデル 2，モデル 3 とした．また，各モデルの隠れユニット数の決定に関しても，実験 1 と同様の方法により決定し，それぞれ $N_2 = 4$，$N_3 = 4$ とした．

3.2.3 変換行列の最適化

学習後のモデル 2 の隠れユニットから得られる系列は，データセット 2 をモデル 2 に入力することにより得た．これらの系列を長さ $T = 752$ の $\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6$ とする．この系列は実験 1 のものと共通である．学習後のモデル 3 の隠れユニットから得られる系列は，データセット 3 をモデル 3 に入力することにより得た．これら四つの系列を長さ $T = 752$ の $\boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}$ とする．

本実験では，等価性構造を適切に抽出するために，目的関数を設定し，関数の最小化を行った．モデル 2 を変換する行列を $\boldsymbol{A}_2$，モデル 3 を変換する行列を $\boldsymbol{A}_3$ とすると，二つの行列は

$$\boldsymbol{A}_2 = \begin{pmatrix} a^{(2)}_{1,1} & a^{(2)}_{2,1} \\ a^{(2)}_{1,2} & a^{(2)}_{2,2} \\ a^{(2)}_{1,3} & a^{(2)}_{2,3} \\ a^{(2)}_{1,4} & a^{(2)}_{2,4} \end{pmatrix}$$

$$\boldsymbol{A}_3 = \begin{pmatrix} a^{(3)}_{1,1} & a^{(3)}_{2,1} \\ a^{(3)}_{1,2} & a^{(3)}_{2,2} \\ a^{(3)}_{1,3} & a^{(3)}_{2,3} \\ a^{(3)}_{1,4} & a^{(3)}_{2,4} \end{pmatrix}$$

と表される．さらに K を変換前の系列数，K' を変換後の系列数とおくと，最小化する目的関数 F は以下の式 8 で表される．

$$F = \sum_{k'=1}^{K'} \sum_{t=1}^{T} \left(\sum_{k=1}^{K} x^{(2)}_{t,k} a^{(2)}_{k,k'} - \sum_{k=1}^{K} x^{(3)}_{t,k} a^{(3)}_{k,k'} \right)^2 \qquad (8)$$

ここで，$a^{(m)}_{k,k'}$ における m はモデル番号である．そして式 8 における $a^{(m)}_{k,k'}$ は各モデル（モデル 2，モデル 3）の隠れ層出力である $\boldsymbol{x}_3, \ldots, \boldsymbol{x}_{10}$ を式 8 を満たすように変換する．そのため，目的関数を最小化するような A_2，A_3 を導出する必要がある．以下に示す Algrorithm1 の操作で目的関数を最小化していく．さらに，F を最小化するにあたって，自明な解である，

$$\boldsymbol{A}_2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\boldsymbol{A}_3 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

が解となることを避けるために，$\boldsymbol{A}_2$，$\boldsymbol{A}_3$ を更新するたびに，$\boldsymbol{A}_2$，$\boldsymbol{A}_3$ の各列の大きさがそれぞれ 1 になるような正規化を行った．このとき，行列 $\boldsymbol{A}$ の上付き文字は行列の列番号を表し，各分数の分母はユークリッドノルムを表している．実際の計算アルゴリズムは以下の Algorithm1 で表される．

本実験では，$\alpha = 0.001$ とした．また，F の値が 0.001 を

Algorithm 1 optimize $\boldsymbol{A}_2$, $\boldsymbol{A}_3$

allocate random values to $\boldsymbol{A}_2$ and $\boldsymbol{A}_3$
compute $\partial F/\partial \boldsymbol{A}_2$ and $\partial F/\partial \boldsymbol{A}_3$
while $F > 10^{-3}$ **do**
  $\boldsymbol{A}_2 \leftarrow \boldsymbol{A}_2 - \alpha \cdot \partial F/\partial \boldsymbol{A}_2$
  $\boldsymbol{A}_3 \leftarrow \boldsymbol{A}_3 - \alpha \cdot \partial F/\partial \boldsymbol{A}_3$
  $\boldsymbol{A}_2^{(1)} \leftarrow \boldsymbol{A}_2^{(1)} / \left|\boldsymbol{A}_2^{(1)}\right|$
  $\boldsymbol{A}_2^{(2)} \leftarrow \boldsymbol{A}_2^{(2)} / \left|\boldsymbol{A}_2^{(2)}\right|$
  $\boldsymbol{A}_3^{(1)} \leftarrow \boldsymbol{A}_3^{(1)} / \left|\boldsymbol{A}_3^{(1)}\right|$
  $\boldsymbol{A}_3^{(2)} \leftarrow \boldsymbol{A}_3^{(2)} / \left|\boldsymbol{A}_3^{(2)}\right|$
  compute F
end while

下回るまで最適化を行った．$\boldsymbol{X}_2 = \langle \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6 \rangle$,$\boldsymbol{X}_3 = \langle \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10} \rangle$ とおく．$\boldsymbol{X}_2$ が変換された後の系列を $\hat{\boldsymbol{X}}_2$，$\boldsymbol{X}_3$ が変換された後の系列を $\hat{\boldsymbol{X}}_3$ とすると，$\hat{\boldsymbol{X}}_2$，$\hat{\boldsymbol{X}}_3$ はそれぞれ

$$\hat{\boldsymbol{X}}_2 = \boldsymbol{X}_2 \cdot \boldsymbol{A}_2 \tag{9}$$

$$\hat{\boldsymbol{X}}_3 = \boldsymbol{X}_3 \cdot \boldsymbol{A}_3 \tag{10}$$

で表すことができる．これらの系列から等価性構造の抽出を試みる．

3.2.4 結果

上述までの変換行列の最適化を経たのちに，等価性構造抽出技術を適用することで等価性構造が得られるか否かを確認した．本実験において，等価性構造抽出に用いたパラメータは，$\tau = 50, \theta_{MSV} = 0.01, \theta_{th} = 0.3$ である．

表 4　抽出された等価性構造（実験 2）

$\boldsymbol{S}_2^{*(2)}$	$v^{(k)}$
	$\langle \#2, \#1 \rangle$
	$\langle \#8, \#7 \rangle$

その結果，モデル 2 における 2-タプル $\langle \#2, \#1 \rangle$ と，モデル 3 の 2-タプル $\langle \#8, \#7 \rangle$ を要素とする等価性構造が抽出された．

3.3 考察

モデル 2 は図形の X 座標，Y 座標の変化，図形の大きさ，図形の形の四つの属性の変化を学習した．モデル 3 は図形の X 座標，Y 座標の変化，図形の大きさ，図形の角度の四つの属性の変化を学習した．今回の実験で，両モデルの隠れ層の出力を変換行列によって回転することにより，共通する属性の変化に反応している隠れユニット群を，等価性構造という形で抽出することができた．しかし今回の実験では，隠れユニットの ID が変換行列によって意味を持たない．この結果から，今回用いたような小規模なニューラルネットワークであれば，互いのデータセットの属性の変化における共通部分に反応しているような隠れユニット群を，ID が意味を持たない形であれば抽出できることがわかった．この実験を拡張すれば，それぞれの MLP の隠れユニットが，属性の変化を適切に認識しているのならば，全く違うタスクに使用される二つのニューラルネットワークの隠れ層の出力から，自明ではない隠れユニット群が共通の振る舞いとして抽出できる可能性があり，転移学習の転移先の同定に適用することも考えられる．

4. まとめと今後の課題

本論文では，ニューラルネットワークの内部表現の理解のために，等価性構造が適用できるかどうかを確かめた．実験 1 では，片方のモデルの属性の変化が，もう片方の属性の変化を内包するようなデータセット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，片方の隠れ層出力に線形変換することによる前処理を加えることで，等価性構造が抽出できることを示した．これにより，隠れ層の出力系列に適切な前処理を加えれば，等価性構造抽出を適用して分析が行える可能性があることがわかった．また

実験 2 では，実験 1 に比べてデータセットをより複雑化し，両モデルの隠れユニットがどのように対応しているのかがわからない状況を扱った．その上で等価性構造を抽出するために，本実験では前処理として，目的関数を設定し，その関数を最小化することで各モデルの隠れ層出力の変換を行った．この実験により，各隠れユニットのおける対応関係が不明な状態でも等価性構造を抽出できることがわかった．

今後の課題としては，入力系列が本論文のものよりも複雑化した場合に，最適化しなければならない変数の数が隠れユニットの数に従って増えるので，今回使用した最急降下法やその他の既知の最適化手法では難しいものも発生することが考えられる．その対策として，新たなる前処理の手法の開発や，前処理を必要としない等価性構造抽出手法の考案などが考えられる．また，二つのデータセットの動きが同期しているため，等価性構造抽出の強みを発揮できる課題とはなっていないが，今後は非同期なデータセットに拡張し，等価群があるかどうかが不明な状況での実験を行うことで，人間がデータセットを見ただけでは見つけることのできないような知見を抽出することが次の課題である．

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．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。—

参考文献

[1] Le, Q. V.: Building high-level features using large scale unsupervised learning, *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, pp. 8595–8598 (2013).

[2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. and Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, Vol. 542, No. 7639, p. 115 (2017).

[3] Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S. et al.: Detecting cancer metastases on gigapixel pathology images, *arXiv preprint arXiv:1703.02442* (2017).

[4] Holden, D., Komura, T. and Saito, J.: Phase-functioned neural networks for character control, *ACM Transactions on Graphics (TOG)*, Vol. 36, No. 4, p. 42 (2017).

[5] 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*, pp. 1–4 (2013).

[6] 立石雅彦，山崎晴明ほか：手書数字認識における階層型ニューラルネットワークの中間層に関する考察，情報処理学会論文誌，Vol. 30, No. 10, pp. 1281–1288 (1989).

[7] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D. and DiCarlo, J. J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex, *Proceedings of the National Academy of Sciences*, Vol. 111, No. 23, pp. 8619–8624 (2014).

[8] Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M. and Cun, Y. L.: Learning convolutional feature hierarchies for visual recognition, *Advances in neural information processing systems*, pp. 1090–1098 (2010).

[9] Zeiler, M. D. and Fergus, R.: Visualizing and understanding convolutional networks, *European conference on computer vision*, Springer, pp. 818–833 (2014).

[10] Minnen, D., Isbell, C., Essa, I. and Starner, T.: Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery, *The IEEE International Conference on Data Mining*, IEEE, pp. 601–606 (2007).

[11] 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, pp. 269–300 (2005).

[12] Lonardi, S., J.Lin, Keogh, E. and Pranav, T.: Finding motifs in time series, *Proc. of the 2nd Workshop on Temporal Data Mining*, pp. 53–68 (2002).

[13] Mueen, A. and Keogh, E.: Online discovery and maintenance of time series motifs, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1089–1098 (2010).

[14] Vahdatpour, A., Amini, N. and Sarrafzadeh, M.: Toward Unsupervised Activity Discovery Using Multi-Dimensional Motif Detection in Time Series., *International Joint Conference on Artificial Intelligence*, Vol. 9, pp. 1261–1266 (2009).

[15] Patel, P., Keogh, E., Lin, J. and Lonardi, S.: Mining motifs in massive time series databases, *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, IEEE, pp. 370–377 (2002).

[16] Hao, M. C., Marwah, M., Janetzko, H., Dayal, U., Keim, D. A., Patnaik, D., Ramakrishnan, N. and Sharma, R. K.: Visual exploration of frequent patterns in multivariate time series, *Information Visualization*, Vol. 11, No. 1, pp. 71–83 (2012).

[17] Hao, Y., Shokoohi-Yekta, M., Papageorgiou, G. and Keogh, E.: Parameter-free audio motif discovery in large data archives, *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, IEEE, pp. 261–270 (2013).

[18] Syed, Z., Stultz, C., Kellis, M., Indyk, P. and Guttag, J.: Motif discovery in physiological datasets: a methodology for inferring predictive elements, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 4, No. 1, p. 2 (2010).

[19] Satoh, S. and Yamakawa, H.: Incremental extraction of high-dimensional equivalence structures, *International Joint Conference on Neural Networks, Anchorage, USA, 2017*, pp. 1518–1524 (online), DOI: 10.1109/IJCNN.2017.7966032 (2017).

[20] Matthey, L., Higgins, I., Hassabis, D. and Lerchner, A.: dSprites:Disentanglement testing Sprites dataset, https://github.com/deepmind/dsprites-dataset/ (2017).