CF-007

多層ニューラルネットにおけるコミュニティ構造の推定と推論におけるコミュニティ間の関係解析

渡邊千紘 †，平松薫 †，柏野邦夫 †

†NTT コミュニケーション科学基礎研究所 〒243-0198 神奈川県厚木市森の里若宮 3-1

概要

多層ニューラルネットは多様な実データに対し高い予測精度を達成しているが，その推論の仕組みはブラックボックス化されており，人間が理解することは難しい．この課題を解決するために，我々はこれまで，ネットワーク解析を用いて多層ニューラルネットにおけるユニットのコミュニティ構造を推定し，元のネットワークを単純化した表現を得る方法を提案し，さらに各コミュニティの役割を入出力マッピングの観点から定量的に求める手法を提案してきた．これらの手法では，学習済みネットワークから抽出された各コミュニティの単独での役割については詳細に知ることが可能となったものの，異なるコミュニティ間の関連性について定量的に知る方法は存在していなかった．本研究では，ニューラルネットの異なる層における任意の 2 つのコミュニティに対し，入力層側のコミュニティから出力層側のコミュニティに与える影響の大きさを定量化し，可視化する手法を提案する．また，実際に回転を含む画像の認識を行うニューラルネットからコミュニティ構造を抽出し，その結果を解析することにより，多くのコミュニティで回転不変な入力画像の特徴に基づく推論が行われていることを確認し，コミュニティ間の関係について考察を行った．

キーワード：多層ニューラルネット，解釈可能性，コミュニティ抽出，画像認識

1 はじめに

深層学習は多くのデータにおいて高い予測性能を実現しており，近年様々な課題に対して用いられるようになってきた．しかし，多層ニューラルネットによるデータの学習結果は非常に多くのパラメータの階層的な組み合わせとして表現されるため，その推論の仕組みはブラックボックス化されており，人間が理解することは難しい．このような**解釈性**の問題は，特に医療など，モデルによる予測結果の可否を人間が判断するような場面においても深層学習を安心して用いることができるようになるために，重要な研究課題となっている．

近年，深層学習における推論の仕組みを解釈することを目的とした研究が，様々なアプローチからなされている．例えば，データから学習されたニューラルネットの働きを，より単純な解釈性の高いモデル (線形モデルや決定木など) で近似することによるアプローチ [1, 2, 3] や，ニューラルネットの予測結果が各入力データ (特に画像データ) のどの部分に影響を受けるかを可視化するアプローチ [4, 5, 6]，ニューラルネットにおけるユニット (集合) の役割やそれらの間の類似度を求めるアプローチ [7, 8, 9]，さらに学習結果が解釈可能な関数になるようなニューラルネットの学習法を構築するアプローチ [10, 11] などが提案されている．これらの手法により，ニューラルネットの全体としての働きや，各部分 (ユニットや層など) の役割など，異なる側面から深層学習の内部の仕組みを捉えることが可能となったが，ニューラルネットの内部構造がどのような部分構造から成り立っており，またその各部分は推論においてどのような役割を果たしているのか，ということに関して自動的に知識を獲得するための手法は存在していなかった．

そこで，我々はこれまで，データから学習されたニューラルネットにおける推論のメカニズムを理解しやすい形に変換するための 1 つのアプローチとして，ニューラルネットにおけるユニットを類似した役割を持つ集合（**コミュニティ**）に分類し，さらに抽出された各コミュニティの推論における役割を解析する手法 [12, 13, 14, 15, 16, 17] を提案してきた．これらの手法は，既存のアプローチとは異なり，学習されたネットワークの全体構造をこれらの手法においては，主に隣接層との結合パターンの情報からユニットのコミュニティ構造を推定し，その後 (1) 隣接するコミュニティ間に存在する結合を閾値処理に基づき単一の結合束で表現する手法 [12, 13, 14, 15, 17] や，(2) 各コミュニティと関連する入出力次元を可視化する手法 [16] を適用することにより，各コミュニティが推論において果たす役割を考察することを可能にしていた．

上記 (1) の手法を用いた場合においては，各コミュニティの役割は結合束により単純化されたニューラルネットの構造から考察可能であるが，各コミュニティ間の関連性の強さについて定量的に知ることはできなかった．また，上記 (2) の手法を用いた場合においては，各コミュニティが入力から受ける影響と出力に与える影響を定量的に知ることができるが，各コミュニティ同士の関連性を可視化する方法は存在していなかった．

そこで，我々は，上記に挙げた手法を用いて学習済みネットワークから抽出されたコミュニティ構造において，ある（入力層側の）コミュニティが異なる（出力層側の）コミュニティに対して与える影響の強さを定量化することで，コミュニティ同士の間の関連性を求めるこ

とを可能にする手法を提案する．また，実際に提案法を用いて回転を含む画像データから学習されたニューラルネットの解析を行うことにより，回転不変な画像特徴を抽出するニューラルネットの内部における推論の構造を可視化し，その結果からニューラルネットの各部分が果たす役割について知識を獲得できることを示す．

2 ニューラルネットの学習とコミュニティ構造推定

提案法は，データから学習されたニューラルネットの構造に対し，後処理としてコミュニティ構造の推定と各コミュニティの役割の解析を行うものであり，全体としては，(a) ニューラルネットの学習，(b) コミュニティ構造の推定，(c) 各コミュニティの役割と異なるコミュニティ間の関係の解析，という流れで実現できる．

本節では，既存研究 [16, 17] の手法に基づき，(a) ニューラルネットの学習，(b) コミュニティ構造の推定，(c) 各コミュニティの入出力マッピングにおける役割の解析を行う方法について述べる．本研究で新たに提案する，(c) の異なるコミュニティ間の関係解析の手法については，次節で述べる．

2.1 (a) ニューラルネットの学習

ニューラルネットの学習法としては，既存研究 [16, 17] と同様，L1 正規化（LASSO [18, 19]）を含めた誤差逆伝播法 [20, 21] に基づいて行うことにより，疎なネットワークを獲得し，過学習を防ぐことができる．詳細な説明は省略するが，各クラスに属する入出力データの組を 1 つずつ順番に選び，そのデータを用いて出力層側から入力層側に向かって各パラメータ（結合重みとバイアス）を更新する処理を繰り返すことにより，ニューラルネットの構造を学習することができる．

2.2 (b) コミュニティ構造の推定

コミュニティ推定においては，学習済みネットワークの各層において，隣接する層との結合パターンの類似性からユニットの分類を行う．ニューラルネットからコミュニティの抽出を行う手法としては様々なものが提案されているが [12, 13, 14, 15, 16, 17] ，本研究では，[16] における手法を用いてコミュニティ推定を行うこととした．この手法は，複雑ネットワークのデータに対しコミュニティ抽出を行う手法 [22] を多層ニューラルネット向けに拡張したものとなっている．

学習済みネットワークのある層に着目したとき，この層と隣接する層との間の結合関係は 4 つの隣接行列 A^+, A^-, B^+, B^- で表現できる．例えば，隣接行列 A^+ は，入力側の隣接層と着目する層の間の正の結合重みによるネットワークを表し，その各要素 $A^+_{i,k}$ は，入力側の隣接層における i 番目のユニットと，着目する層における k 番目のユニットとの間に ξ 以上の結合重みが存在していれば $A^+_{i,k}=1$, そうでなければ $A^+_{i,k}=0$ と定義できる．ここで，ξ は正の値をとるハイパーパラメータである．同様に，隣接行列 A^- は，入力側の隣接層における i 番目のユニットと，着目する層における k 番目のユニットとの間に $-\xi$ 以下の結合重みが存在していれば $A^-_{i,k}=1$, そうでなければ $A^-_{i,k}=0$ と定義できる．隣接行列 B^+, B^- に関しても，着目する層における k 番目のユニットと，出力側の隣接層における j 番目のユニットとの間の結合重みから同様に各要素 $B^+_{k,j}$, $B^-_{k,j}$ が定義できる．

上記の 4 つの隣接行列を観測データとして，以下のような確率モデルを仮定する．まず，着目する層において，あるユニットがコミュニティ c に属する確率をパラメータ π_c で表す．ただし，π_c は以下を満たすものとする：$\sum_c \pi_c = 1$. 次に，着目する層におけるコミュニティ c に属するユニットと，入力層側の隣接層における i 番目のユニットとの間に正負の結合が存在する確率を，それぞれパラメータ $\tau^+_{i,c}$, $\tau^-_{i,c}$ で表す．同様に，着目する層におけるコミュニティ c に属するユニットと，出力層側の隣接層における j 番目のユニットとの間に正負の結合が存在する確率を，それぞれパラメータ $\tau'^+_{k,j}$, $\tau'^-_{k,j}$ で表す．着目する層において，各ユニット k が属するコミュニティを g_k とすると，上記のパラメータが与えられたときの隣接行列 A^+, A^-, B^+, B^- と $g=\{g_k\}$ の確率は以下で与えられる．

$$\begin{aligned}&\Pr(A^+,A^-,B^+,B^-,g|\pi,\tau^+,\tau^-,\tau'^+,\tau'^-)\\&=\Pr(A^+,A^-,B^+,B^-|g,\pi,\tau^+,\tau^-,\tau'^+,\tau'^-)\\&\Pr(g|\pi,\tau^+,\tau^-,\tau'^+,\tau'^-).\end{aligned}$$

ただし，

$$\begin{aligned}&\Pr(A^+,A^-,B^+,B^-|g,\pi,\tau^+,\tau^-,\tau'^+,\tau'^-)\\&=\prod_k\Big\{\prod_i\left(\tau^+_{g_k,i}\right)^{A^+_{i,k}}\left(1-\tau^+_{g_k,i}\right)^{1-A^+_{i,k}}\left(\tau^-_{g_k,i}\right)^{A^-_{i,k}}\\&\left(1-\tau^-_{g_k,i}\right)^{1-A^-_{i,k}}\Big\}\Big\{\prod_j\left(\tau'^+_{g_k,j}\right)^{B^+_{k,j}}\left(1-\tau'^+_{g_k,j}\right)^{1-B^+_{k,j}}\\&\left(\tau'^-_{g_k,j}\right)^{B^-_{k,j}}\left(1-\tau'^-_{g_k,j}\right)^{1-B^-_{k,j}}\Big\},\end{aligned}$$

$$\Pr(g|\pi,\tau^+,\tau^-,\tau'^+,\tau'^-)=\prod_k\pi_{g_k}. \tag{1}$$

とした．

各ユニット k に対するコミュニティ割り当て $g=\{g_k\}$ は隠れ変数であるため，これについて対数尤度の期待値 $\bar{\mathcal{L}}$ を取ると

$$\begin{aligned}\bar{\mathcal{L}}=\sum_{k,c}q_{k,c}\Big\{&\ln\pi_c+\sum_i(A^+_{i,k}\ln\tau^+_{c,i}+(1-A^+_{i,k})\\&\ln(1-\tau^+_{c,i})+A^-_{i,k}\ln\tau^-_{c,i}+(1-A^-_{i,k})\ln(1-\tau^-_{c,i}))\\&+\sum_j(B^+_{k,j}\ln\tau'^+_{c,j}+(1-B^+_{k,j})\ln(1-\tau'^+_{c,j})\\&+B^-_{k,j}\ln\tau'^-_{c,j}+(1-B^-_{k,j})\ln(1-\tau'^-_{c,j}))\Big\},\end{aligned}$$

ここで，着目する層においてユニット k がコミュニティ c に属する確率を $q_{k,c}$ とおいた．

$$q_{k,c} \equiv \Pr(g_k = c|A^+, A^-, B^+, B^-, \pi, \tau^+, \tau^-, \tau'^+, \tau'^-) = \frac{\Pr(A^+, A^-, B^+, B^-, g_k = c|\pi, \tau^+, \tau^-, \tau'^+, \tau'^-)}{\Pr(A^+, A^-, B^+, B^-|\pi, \tau^+, \tau^-, \tau'^+, \tau'^-)}. \quad (2)$$

証明は省略するが（参考文献 [16] を参照），上記の対数尤度の期待値 $\bar{\mathcal{L}}$ を最大化する最適なパラメータ $\pi, \tau^+, \tau^-, \tau'^+, \tau'^-$ と，そのときの $q_{k,c}$ は，以下の式を満たす．

$$q_{k,c} = \frac{r_{k,c}}{\sum_s r_{k,s}}, \quad (3)$$

$$\pi_c = \frac{\sum_k q_{k,c}}{k_0},\ \tau^+_{c,i} = \frac{\sum_k A^+_{i,k} q_{k,c}}{\sum_k q_{k,c}},\ \tau^-_{c,i} = \frac{\sum_k A^-_{i,k} q_{k,c}}{\sum_k q_{k,c}},$$
$$\tau'^+_{c,j} = \frac{\sum_k B^+_{k,j} q_{k,c}}{\sum_k q_{k,c}},\ \tau'^-_{c,j} = \frac{\sum_k B^-_{k,j} q_{k,c}}{\sum_k q_{k,c}}. \quad (4)$$

ここで，k_0 を着目する層におけるユニット数とし，

$$r_{k,c} \equiv \pi_c \left[\prod_i \left(\tau^+_{c,i}\right)^{A^+_{i,k}} \left(1-\tau^+_{c,i}\right)^{1-A^+_{i,k}} \left(\tau^-_{c,i}\right)^{A^-_{i,k}} \left(1-\tau^-_{c,i}\right)^{1-A^-_{i,k}}\right] \left[\prod_j \left(\tau'^+_{c,j}\right)^{B^+_{k,j}} \left(1-\tau'^+_{c,j}\right)^{1-B^+_{k,j}} \left(\tau'^-_{c,j}\right)^{B^-_{k,j}} \left(1-\tau'^-_{c,j}\right)^{1-B^-_{k,j}}\right].$$

とおいた．上記の式 (3) と式 (4) に従い，パラメータ $\pi, \tau^+, \tau^-, \tau'^+, \tau'^-$ と $\{q_{k,c}\}$ を交互に更新することにより，局所的に最適な解を得ることができる．この更新処理を反復し，最終的に得られた $\{q_{k,c}\}$ の値に基づき，各ユニット k の属するコミュニティの推定結果を $\arg\max_c q_{k,c}$ として得ることができる．

2.3 (c) 各コミュニティの入出力マッピングにおける役割の解析

上記の手法に基づき，ニューラルネットから抽出された各コミュニティが，予測においてどのような役割を担っているのかを定量的に解析するための手法が提案されている [16]．これは，各コミュニティ c について，各入力次元 i の値から受ける影響の大きさと，各出力次元 j の値に与える影響の大きさを測り，それぞれの値を並べて特徴ベクトルとすることで実現することができる．ここで，各入力次元 i からコミュニティ c が受ける影響を表す特徴ベクトルを $v^{\mathrm{in}}_c = \{v^{\mathrm{in}}_{ic}\}$ とし，コミュニティ c から各出力次元 j が受ける影響を表す特徴ベクトルを $v^{\mathrm{out}}_c = \{v^{\mathrm{out}}_{cj}\}$ とする．以下に，これらの特徴ベクトルを求めるための手法の詳細を示す．

まず，各入力次元 i からコミュニティ c が受ける影響の大きさ v^{in}_{ic} は，入力データの次元 i の情報が使えなくなった時に，コミュニティ c に含まれるユニットの出力で生じる誤差として定義することができる．より具体的には，以下のように定義する．入力層以外の層において，n 番目の入力サンプル $X^{(n)}$ に対するユニット k の出力を $o^{(n)}_k$ とおく．また，以下のように i 番目の次元の値のみを変更した入力サンプル $X'^{(n)}$ に対するユニット k の出力を $z^{(n)}_k$ とおく．

$$X'^{(n)}_i \equiv \frac{1}{n_1}\sum_n X^{(n)}_i.$$
$$\text{For } l \neq i,\ X'^{(n)}_l \equiv X^{(n)}_l.$$

$u(c)$ をコミュニティ c に含まれる全てのユニットからなる集合として，入力次元 i からコミュニティ c が受ける影響の大きさ v^{in}_{ic} を以下で定義する．

$$v^{\mathrm{in}}_{ic} = \sqrt{\frac{1}{n_1}\sum_{k\in u(c)}\sum_n \left(o^{(n)}_k - z^{(n)}_k\right)^2}.$$

これは，入力データの次元 i の値が（データによらず）学習データに対する平均値で置き換えられた時の，コミュニティ c に含まれるユニットの出力における二乗平均平方根誤差を表している．

同様に，コミュニティ c から各出力次元 j が受ける影響の大きさ v^{out}_{cj} は，コミュニティ c に含まれるユニットの情報が使えなくなった時に，j 番目の出力次元で生じる誤差として定義することができる．より具体的には，以下のように定義する．n 番目の入力サンプル $X^{(n)}$ に対する j 番目の出力次元の値を $y^{(n)}_j$ とおく．また，n 番目の入力サンプルに対し，以下のように（出力層以外の層における）コミュニティ c に含まれるユニット k の出力値を $o^{(n)}_k$ から $o'^{(n)}_k$ に変更したときの，j 番目の出力次元の値を $z^{(n)}_j$ とおく．

$$\text{For } k \in u(c),\ o'^{(n)}_k \equiv \frac{1}{n_1}\sum_n o^{(n)}_k.$$
$$\text{For } k \notin u(c),\ o'^{(n)}_k \equiv o^{(n)}_k.$$

コミュニティ c から各出力次元 j が受ける影響の大きさ v^{out}_{cj} を以下で定義する．

$$v^{\mathrm{out}}_{cj} = \sqrt{\frac{1}{n_1}\sum_n \left(y^{(n)}_j - z^{(n)}_j\right)^2}.$$

これは，コミュニティ c に含まれるユニットの出力が（データによらず）学習データに対する平均値で置き換えられた時の，出力次元 j における二乗平均平方根誤差を表している．

このようにして獲得した 2 種類の特徴ベクトル v^{in}_c, v^{out}_c を基にして，各コミュニティ c が入出力マッピングにおいて果たす役割を考察することができる．

3 多層ニューラルネットにおける異なるコミュニティ間の関係解析

2 節に述べた一連の手法では，学習済みネットワークをコミュニティ構造に分割し，さらに各コミュニティについてその推論における単独での役割を定量化することが可能となったが，異なるコミュニティ間の関係について知識を得るための方法は存在していなかった．

そこで，我々は，ニューラルネットから抽出されたコミュニティ構造における任意の異なる 2 つのコミュニティについて，それらの間の関連性を定量的に評価する手法を提案する．これは，2.3 節に述べた各コミュニティの役割の定量化法を拡張し，任意の入力層側のコミュニティ c に含まれるユニットの出力値が，出力層側のコミュニティ c' に含まれるユニットの出力値に対して与える影響の大きさ $s = \{s_{cc'}\}$ を解析することで実現できる．以下に，入力層側のコミュニティ c が出力層側のコミュニティ c' に対して与える影響の大きさを定量化するための手法の詳細を示す．

$s_{cc'}$ は，入力層側のコミュニティ c の情報が使えなくなった時に，コミュニティ c' に含まれるユニットの出力で生じる誤差として定義することができる．より具体的には，以下のように定義する．n 番目の入力サンプル $X^{(n)}$ に対するユニット k' の出力を $o_{k'}^{(n)}$ とおく．また，入力サンプル $X^{(n)}$ に対し，以下のようにコミュニティ c に含まれるユニット k の出力値を $o_k^{(n)}$ から $o_k'^{(n)}$ に変更したときの，ユニット k' の出力を $z_{k'}^{(n)}$ とおく．

$$\text{For } k \in u(c),\ o_k'^{(n)} \equiv \frac{1}{n_1}\sum_n o_k^{(n)}.$$

$$\text{For } k \notin u(c),\ o_k'^{(n)} \equiv o_k^{(n)}.$$

このとき，$s_{cc'}$ を以下で定義する．

$$s_{cc'} = \sqrt{\frac{1}{n_1}\sum_{k' \in u(c')}\sum_n \left(o_{k'}^{(n)} - z_{k'}^{(n)}\right)^2}.$$

これは，コミュニティ c に含まれるユニットの出力が（データによらず）学習データに対する平均値で置き換えられた時の，コミュニティ c' に含まれるユニットの出力における二乗平均平方根誤差を表している．このようにして獲得した行列 $s = \{s_{cc'}\}$ から，任意のコミュニティ対 (c, c') について，コミュニティ c がコミュニティ c' に与える影響の大きさを知ることができる．

4 実験

2, 3 節に述べた手法を実際のデータに対して適用し，ニューラルネットの学習結果から得られる知識に関して考察する．特に，本研究では，0°, 90°, 180°, 270° の回転を含む 10 クラスの図形画像 (20×20 画素) を分類するという課題のもと，ニューラルネットを学習し，コミュニティ構造を推定した後，各コミュニティの単独での役割と異なるコミュニティ間の関係を解析する．

詳細な実験条件を以下に示す．まず，既存研究 [16] におけるデータと同じ生成方法を用い，0°, 90°, 180°, 270° の回転を加えることにより，回転を含む図形画像のデータを生成した．図 1 に，入力画像のサンプルを示す．これらのデータを用いてニューラルネットを学習する．ここで，学習データ数を 1000, テストデータ数を 1000 とし，1 つの学習データ (入出力データの対) に対する平均の反復数を 100 回とした．入出力データの正規化法，学習時のステップ幅に関しては，既存研究 [16] と同じ設定を用いた．また，LASSO[18, 19] のハイパーパラメータを $\lambda = 1.1 * 10^{-5}$ とし，学習結果として疎なネットワークが得られるようにした上で，絶対値が 0.005 未満の結合重みを削除したネットワーク構造に対してコミュニティ抽出を行った．ここで，1 層のコミュニティ抽出において，コミュニティ数を 10, EM アルゴリズムの反復数を 200 とし，ランダムな初期値に対し 300 回コミュニティ抽出を行った上で，最終反復における対数尤度の期待値が最大となった回の結果を用いることとした．最後に，抽出されたコミュニティ構造に対し，各コミュニティの単独での役割と異なるコミュニティ間の関係について 2, 3 節に述べた手法を用いて解析した．

図 2 にデータから学習されたニューラルネットを，図 3 にニューラルネットから抽出されたコミュニティ構造を示す．図の下側が入力層，上側が出力層に対応する．ただし，図 3 下は，入力層の各コミュニティに含まれる画素の組み合わせを表す．また，図 4 に各コミュニティが入力から受ける影響と出力に与える影響の可視化結果を示す．各層において，Com 1, Com 2, ⋯, Com 10 は，図 3 における左から 1 番目，2 番目，⋯，10 番目のコミュニティと対応しており，各出力次元の値に最も影響を与える（各クラスの図形の分類に最も影響を与える）コミュニティの結果を黒のバーで示した．

図 3 下より，入力層における各コミュニティにおいて，0°, 90°, 180°, 270° の回転に対して対称な形の組み合わせで画素の集合が抽出されており，これらの各コミュニティ内の画素は隠れ層に対し類似した結合パターンを持つようにネットワークが学習されたことが分かる．

各コミュニティの単独での役割について，図 4 の結果から，各層における多くのコミュニティで入力画像の回転対称な特徴が推論に用いられていることが分かる．また，例えば入力層と，隠れ層 2(出力層側の隠れ層) において，長方形 (Rectangle) とダイヤ型 (Diamond) の認識に最も影響を与えるコミュニティは同一のものであり，これらの図形の認識には類似した入力情報が用いられていることが推察される．入力層，隠れ層 1(入力層側の隠れ層) において，多くの図形の認識結果に影響を与える重要なコミュニティ (入力層では Com 10, 隠れ層 1 では Com 2 など) が少数存在している一方で，どの図形

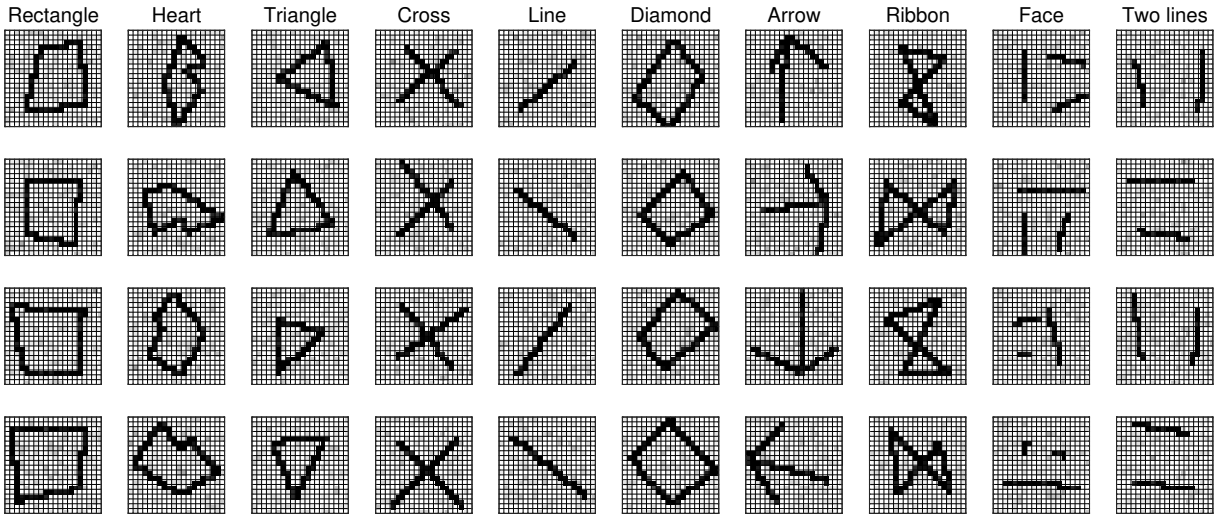

図 1. 0°, 90°, 180°, 270° の回転を含む 10 クラスの入力画像のサンプル.

の認識結果にも比較的影響が少ないコミュニティが多数存在していることが分かる.

さらに，図 5 に，異なるコミュニティ間の関係解析の結果を示す．図 5 が 3 節の手法を用いて獲得された行列 $s = \{s_{cc'}\}$ であり，図 5 中央，右はそれぞれ，行列 s の要素を行方向，列方向の最大値が等しくなるように正規化し，その最大値を取る要素に×印を記入した結果である．図 5 中央の結果から，入力側のコミュニティ (行) を固定したときに，どの出力側のコミュニティに最も影響を与えるかを知ることができ，また図 5 右の結果から，出力側のコミュニティ (列) を固定したときに，どの入力側のコミュニティから最も影響を受けるかを知ることができる.

図 5 中央の結果から，出力層以外における各コミュニティが最も影響を与えるコミュニティは，入力層における Com 10 以外で隣接層のコミュニティであることが分かる．一方，図 5 右の結果から，入力層以外における各コミュニティが最も影響を受けるコミュニティは，層によらず入力層のコミュニティ (特に，入力層の Com 10) である場合が多いことが分かる．出力側のコミュニティから見て最も影響を受けやすい入力側のコミュニティとしては，入力層の Com 2, 8, 10, 隠れ層 1 の Com 2 のみが選ばれている．入力層のコミュニティ抽出結果 (図 3 下) から，Com 10 は入力画像の中心部に位置する画素の情報を表しており，この部分の情報がニューラルネットの多くの部分で用いられていることが分かる．入力層における Com 2, 8 はそれぞれ，図形の角にあたる四隅の部分の情報を表しており，これらの情報も多くのコミュニティで用いられていることが分かる.

5 考察

提案法により，学習されたニューラルネットを構成する各コミュニティについて，推論における単独での役割に加え，異なるコミュニティ間の関連度を定量化することができるようになった．しかしながら，各コミュニティが各入出力次元もしくは別のコミュニティに対して「どのように」影響を与えている/ 与えられているかという情報に関しては，未だ解明するための手法が存在していない．各コミュニティが入力層側に存在する各部分の値をどのように変換して推論に用いているのか，また各コミュニティに含まれるユニットの出力値はどのように変換されて出力層側に存在する各部分の値に影響を与えているのかを知るための手法を構築することは，今後の課題である.

また，提案法を適用したときに得られる結果は，コミュニティ抽出における各層のコミュニティ数や，学習時のハイパーパラメータなどの設定に依存する．これらの設定を，ニューラルネットの汎化性能と解析結果の解釈性の両観点から最適化するための手法を作ることも，重要な課題である.

さらに，本研究では回転を含む画像の認識を行うためのニューラルネットに焦点を当てたが，回転以外にも様々な性質を与えたデータセットに対し，学習されたネットワークの内部構造を解析することにより，データセットの性質に応じてニューラルネットの各部分がどのような特徴を抽出するように学習されるのかということについて知識を獲得できると考えられる.

6 結論

多層ニューラルネットは，既に画像処理や音声認識など多くのタスクにおいて高い予測精度を実現することが

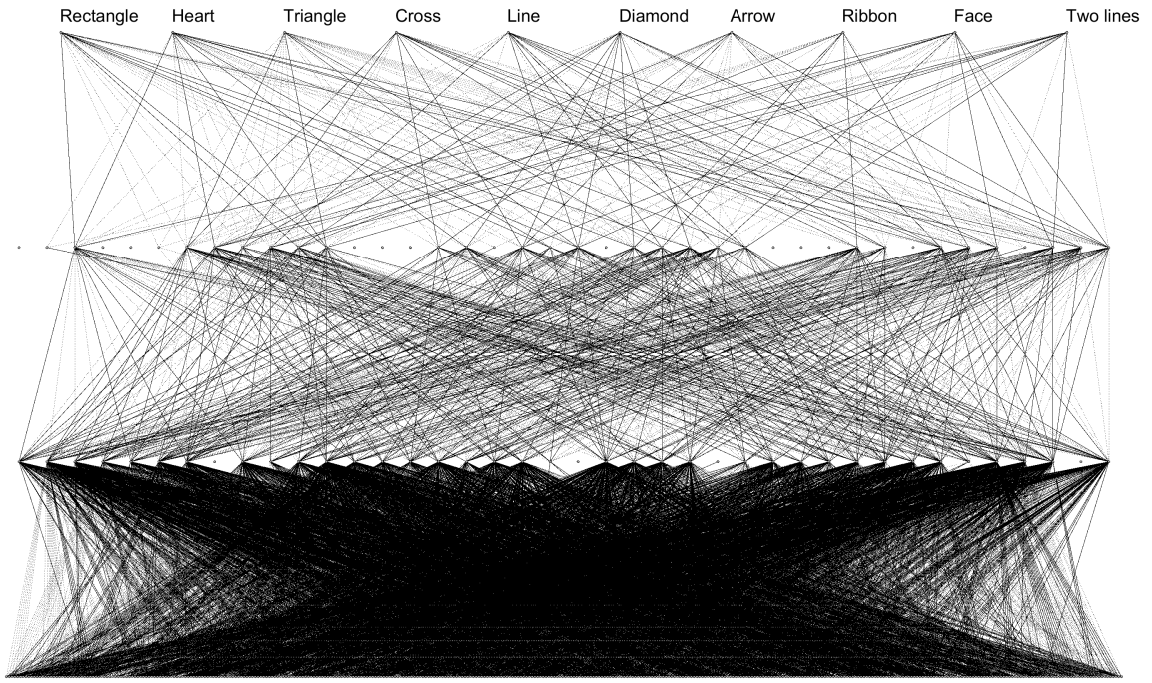

図 2. 図形画像データから学習されたニューラルネットの構造.

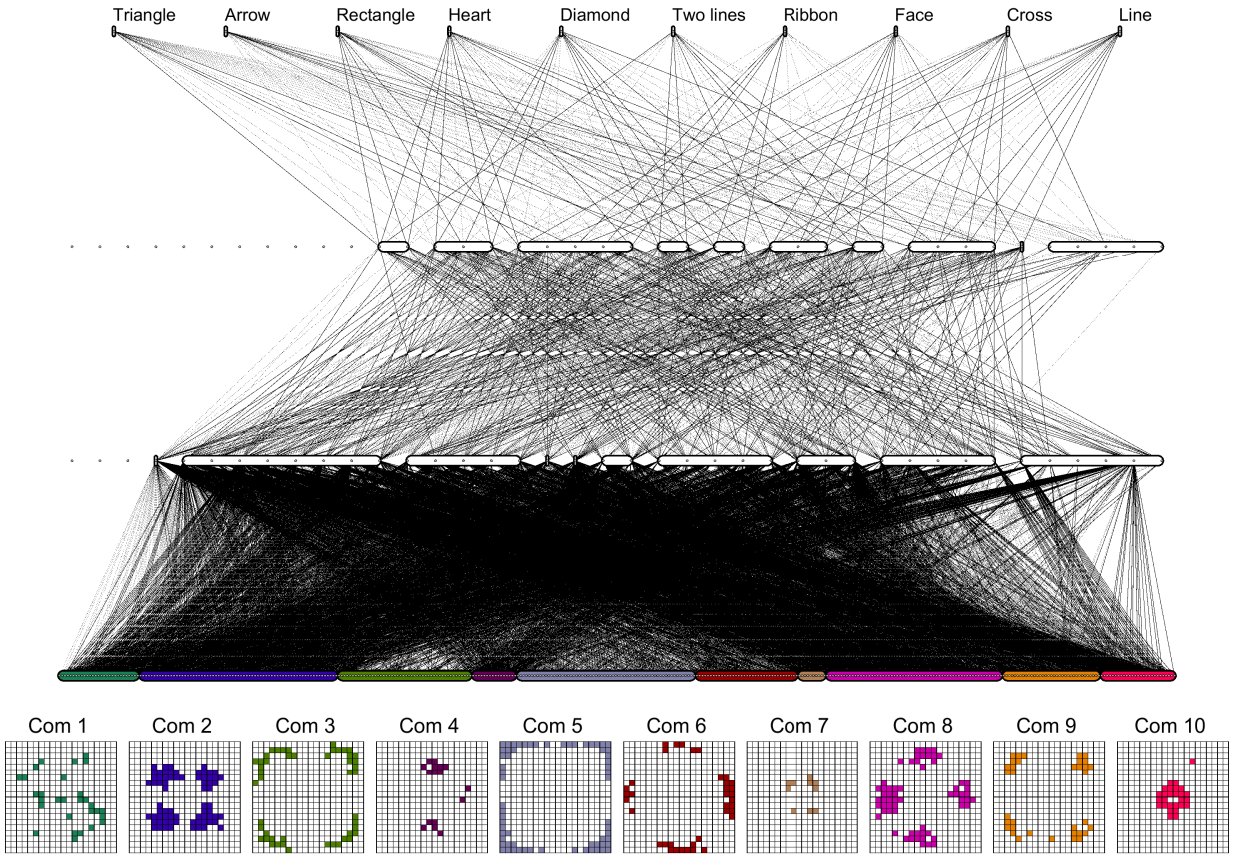

図 3. 多層ニューラルネットから抽出されたコミュニティ構造. 下の図は入力層における各コミュニティに含まれる画素の集合を表す.

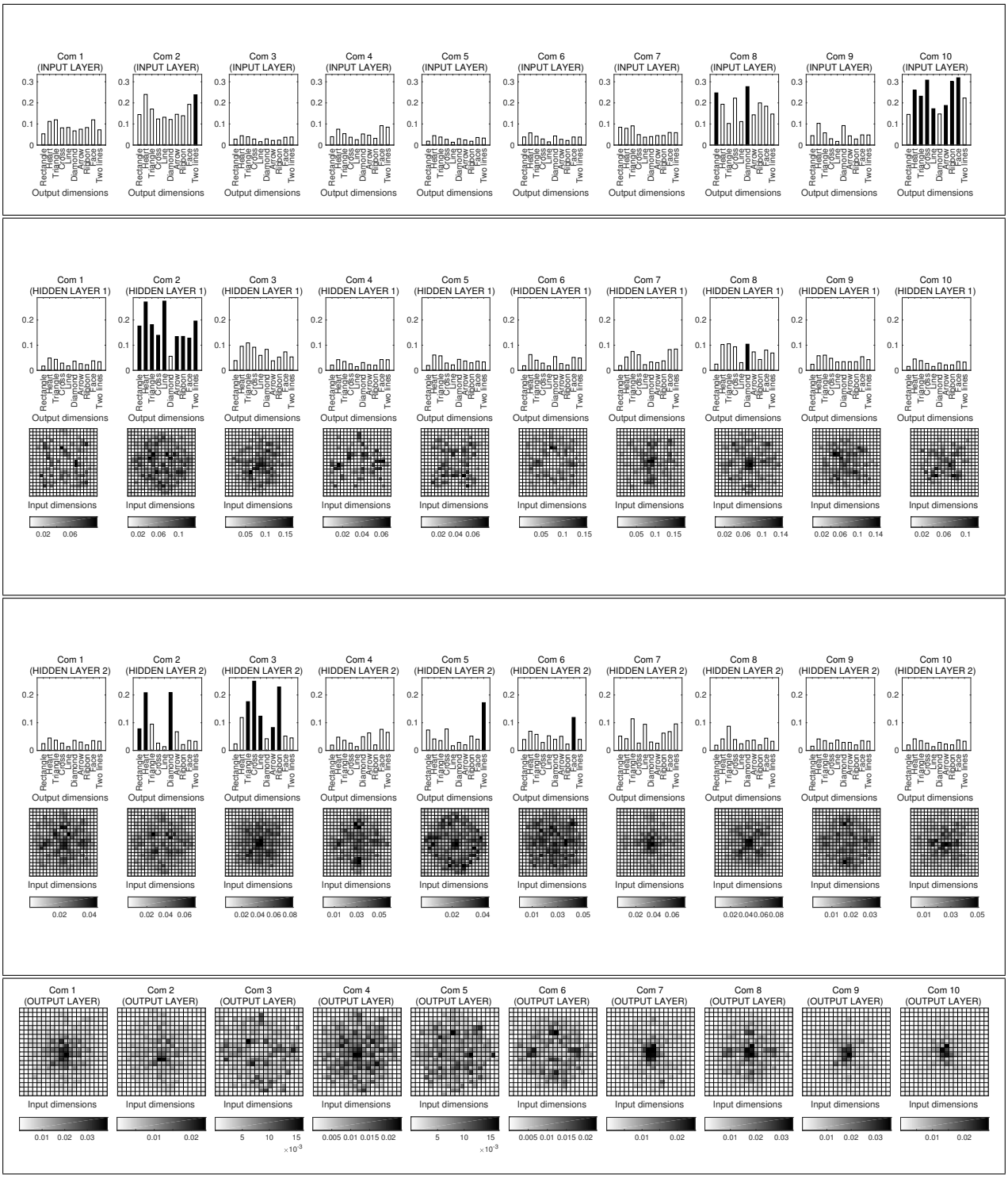

図 4. 各コミュニティが入力から受ける影響と出力に与える影響の可視化結果．各層において，各出力次元の値に最も影響を与える（各クラスの図形の分類に最も影響を与える）コミュニティにおける結果を黒のバーで示した．

確認されており，有用性が示されている一方で，その内部の仕組みを人間が解釈することは困難である．我々はこれまで，データから学習されたニューラルネットの内部構造について知識を獲得するための手法として，学習済みネットワークからコミュニティ構造を抽出し，得られた各コミュニティごとに入出力との関係性から推論における役割を定量的に求めるアプローチを提案してきた．本研究では，さらに，ニューラルネットの異なる層における任意の 2 つのコミュニティについて，入力層側のコミュニティから出力層側のコミュニティに与える影響の大きさを定量化する手法を提案した．また，実際に回転を含む画像の認識を行うニューラルネットに提案法

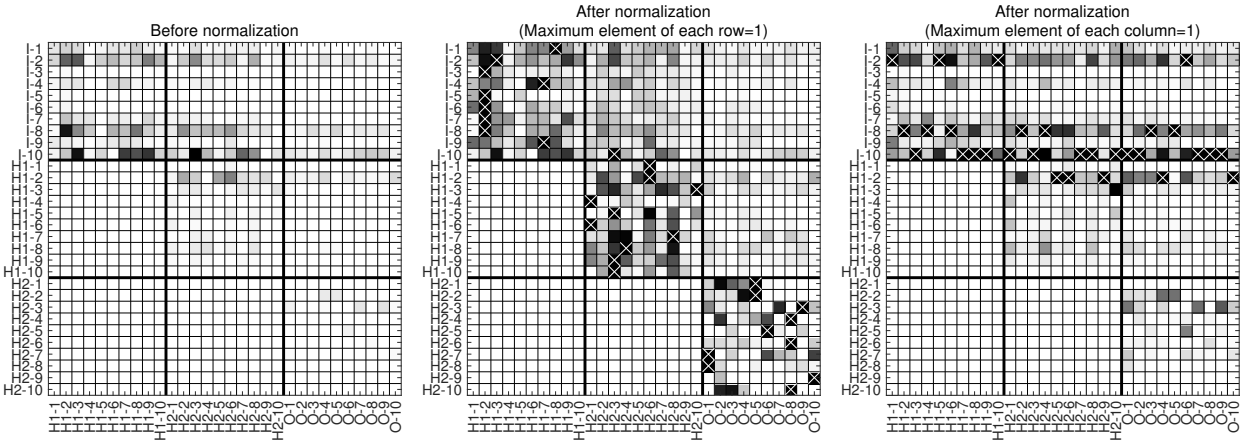

図 5. 左：異なるコミュニティ間の関係の強さを表した行列．中央：左図の行列を，各行について最大値が 1 となるように正規化した行列．×印は各行で最大値を取る要素を表す．右：左図の行列を，各列について最大値が 1 となるように正規化した行列．×印は各列で最大値を取る要素を表す．“I”は入力層，“H1”は入力層に隣接する隠れ層，“H2”は出力層に隣接する隠れ層，“O”は出力層を表し，各数字はコミュニティの番号（図 3, 4 と対応）を表す．

を適用した結果に基づき，各コミュニティの役割と異なるコミュニティ間の関係について考察を行った．

参考文献

[1] M. T. Ribeiro, S. Singh, and C. Guestrin. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[2] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pp. 4765–4774, 2017.

[3] T. Nagamine and N. Mesgarani. Understanding the representation and computation of multilayer perceptrons: A case study in speech recognition. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2564–2573, 2017.

[4] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3145–3153, 2017.

[5] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *6th International Conference on Learning Representations*, 2018.

[6] S. Hara, K. Ikeno, T. Soma, and T. Maehara. Maximally invariant data perturbation as explanation. arXiv:1806.07004, 2018.

[7] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems 29*, pp. 4898–4906, 2016.

[8] T. Zahavy, N. Ben-Zrihem, and S. Mannor. Graying the black box: Understanding DQNs. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1899–1908, 2016.

[9] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pp. 6076–6085, 2017.

[10] J. N. Foerster, J. Gilmer, J. Sohl-Dickstein, J. Chorowski, and D. Sussillo. Input switched affine networks: An RNN architecture designed for interpretability. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1136–1145, 2017.

[11] C. González, E. L. Mencía, and J. Fürnkranz. Re-training deep neural networks to facilitate Boolean concept extraction. In *Proceedings of Discovery Science 2017, Lecture Notes in Computer Science*, Vol. 10558, pp. 127–143, 2017.

[12] C. Watanabe, K. Hiramatsu, and K. Kashino. Modular representation of layered neural networks. *Neural Networks*, Vol. 97, pp. 62–73, 2018.

[13] 渡邊千紘, 平松薫, 柏野邦夫. 多層ニューラルネットにおける正負の結合重みに基づく大局構造抽出. 情報科学技術フォーラム (FIT2017), 2017.

[14] C. Watanabe, K. Hiramatsu, and K. Kashino. Recursive extraction of modular structure from layered neural networks using variational Bayes method. In *Proceedings of Discovery Science 2017, Lecture Notes in Computer Science*, Vol. 10558, pp. 207–222, 2017.

[15] C. Watanabe, K. Hiramatsu, and K. Kashino. Modular representation of autoencoder networks. In *Proceedings of 2017 IEEE Symposium on Deep Learning, 2017 IEEE Symposium Series on Computational Intelligence*, 2017.

[16] C. Watanabe, K. Hiramatsu, and K. Kashino. Understanding community structure in layered neural networks. arXiv:1804.04778, 2018.

[17] 渡邊千紘, 平松薫, 柏野邦夫. 図形認識のための多層ニューラルネットにおける大局構造の抽出. 2018 年度 人工知能学会全国大会 (第 32 回)(JSAI2018), 2018.

[18] M. Ishikawa. A structural connectionist learning algorithm with forgetting. *Journal of Japanese Society for Artificial Intelligence*, Vol. 5, pp. 595–603, 1990.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, Vol. 58, No. 1, pp. 267–288, 1996.

[20] P. Werbos. *Beyond regression : new tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University, 1974.

[21] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, Vol. 323, pp. 533–536, 1986.

[22] J. Wang and C-H. Lai. Detecting groups of similar components in complex networks. *New Journal of Physics*, Vol. 10, No. 123023, 2008.