

クロスエントロピー最小化に基づくネットワークデータの埋め込み

山田 武士[†] 斉藤 和己[†] 上田 修功[†]

ネットワークで表現されたデータを低次元ユークリッド空間へ埋め込む新たな方法を提案する。本手法では、ネットワークの隣接行列が定義するノード間の離散類似度と、現在の埋め込み配置から導かれる連続類似度との間のクロスエントロピーを最小にすることによって最適な配置を求める。さらに、与えられた埋め込みにおいて、ノードの接続関係がいかに忠実に再現されているかを定量的に評価する新たな評価尺度を提案する。最後に実際のネットワークデータを用いた実験によって、本手法の有効性を検証した。

Embedding Network Data based on Cross-Entropy Minimization

TAKESHI YAMADA,[†] KAZUMI SAITO[†] and NAONORI UEDA[†]

We present a novel approach to embed network data into a low dimensional Euclidean space. The proposed method attempts to minimize cross-entropy between the discrete node similarity measure induced by the adjacency matrix and the continuous node similarity function derived from current embedded node positions. We also propose a natural criterion to effectively evaluate an embedded network layout in terms of how well node connectivities are preserved. Experiments using real network data show the effectiveness of the proposed method.

1. ま え が き

複雑なリレーショナルデータの表現手段としてネットワーク、もしくはグラフは様々な研究分野で用いられている。たとえば対象が Web の場合、各 Web ページをノード、ページ間のハイパーリンクをノード間のリンクとするハイパーリンクネットワークがよく用いられるし、生物学では、遺伝子やたんぱく質、代謝物質などの相互作用を遺伝子制御ネットワークとして表現する¹⁾。また人間同士、あるいは企業などの社会的存在同士の関係の解析にはソーシャルネットワークが用いられる²⁾。ネットワーク表現は、グラフ理論などを駆使した数学的な解析を可能にすると同時に、低次元空間に埋め込んで実際に表示させることで、内在するデータの構造的特徴が浮彫りとなり、その結果、研究者にとって新たな科学的発見のための重要な手がかりを与える。このようにネットワークの効果的な埋め込み法の研究は、発見科学および機械学習の立場からみてもきわめて重要である。

ネットワークの構造を直感的に理解し解析する基本

手段の1つとして「ブラウジング」があげられる。ブラウジングとは、ネットワークを低次元の空間に埋め込み、その配置上でノードからノードへのリンクを実際にたどったり、ノードごとの接続関係を相互に比較したりするなど、低次元の配置を直接調べることである。本研究の目的は、このブラウジングに最適なネットワークの低次元ユークリッド空間への埋め込みアルゴリズムを提案することである。

さらに、実際に低次元への埋め込みが得られたとしても、それがブラウジングに適した最適な配置であるかどうかを評価するのは難しい。従来は「美観上好ましい」(aesthetically pleasing)かどうか、などで評価されていたが^{3),4)}、このような尺度は主観的であり、定量的な評価が難しい。

また、これまでの研究の多くは埋め込みの配置を決めるにあたって任意のノード間の距離の存在を仮定している。たとえば Kamada らによって提案されたいわゆるバネモデル⁵⁾(以降、KK バネモデルと呼ぶ)では、任意のノード間のグラフ距離(graph-theoretic distance)を計算し、この距離を長さとするばねをノード間に設定することによって、この距離を最も忠実に保存するような低次元への埋め込みを行う。ここでグラフ距離とは、2ノード間のグラフ上の最短経路長

[†] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

(すなわち最小リンク数) のことであり, Floyd アルゴリズム⁶⁾ などによって計算される. 同様に, 後述する Torgerson による古典的な多次元尺度法 (古典 MDS)⁷⁾ やスペクトラルクラスタリング⁸⁾ などもグラフ距離に基づく手法である. 換言するとこれら従来法は次の原則に基づく手法と考えられる.

[原則DP] (distance preserving principle)

各ノード間のグラフ距離が低次元ユークリッド空間において最も忠実に復元されるように配置する. 一方, 本論文では次のような単純かつ基本的な原則を出発点とする.

[原則CP] (connectivity preserving principle)

各ノードにとって, それと隣接する (すなわち, 直接リンクでつながっている) ノードが, 隣接しないノードより相対的により近くなるように配置する.

原則DP が完全に満足されると **原則CP** も自動的に満足されるが, 逆は成り立たない. また後述するように, 特に埋め込みの次元が低く, かつ大規模で複雑なネットワークの場合, 従来法では **原則DP** を満足することが難しくなり, その結果 **原則CP** も満たされず, ブラウジングに不向きな片寄った配置となることが多い. 一方, **原則CP** に基づいてノード間の隣接関係から直接配置を求めることができれば, 埋め込み先の空間をより有効に活用することによってよりブラウジングに適した配置を得ることができる. したがって本論文では **原則CP** に基づき, ノード間の隣接関係のみに着目したクロスエントロピーを目的関数とするアルゴリズムを提案する.

2. クロスエントロピー最小化に基づく方法

2.1 目的関数

今, N 個のノード (あるいは頂点) からなるネットワーク (あるいはグラフ) を考え, その隣接行列を $A = (a_{i,j})$ とする. 本論文では簡単のため無向ネットワーク (グラフ) のみ扱うこととし, $a_{i,j} \in \{0, 1\}$, $a_{i,i} = 1$ かつ $a_{i,j} = a_{j,i}$ と仮定する. ただし, 本手法は有向ネットワーク (グラフ) にも容易に拡張できる. 与えられたネットワークに対し我々の目的は, K 次元ユークリッド距離の意味で **原則CP** を実現する N 個のノードの K 次元空間への埋め込みを求めることである.

K 次元空間における N 個のノードの座標を $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ とする. ここで $\mathbf{x}_i$ および $\mathbf{x}_j$ 間のユークリッド距離は次式で定義される.

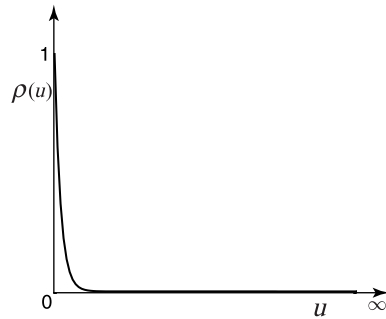


図 1 類似度関数の例

Fig. 1 An example of the similarity function.

$$d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \sum_{k=1}^K (x_{i,k} - x_{j,k})^2 \quad (1)$$

次に $u \geq 0$ に関する単調減少関数 $\rho(u) \in [0, 1]$ を考え, $\rho(0) = 1$, $\rho(\infty) = 0$ とする. このとき $\rho(d_{i,j})$ は $\mathbf{x}_i, \mathbf{x}_j$ 間の (連続値をとる) 類似度関数と見なすことができる. ここで ρ として図 1 に示すような $u = 0$ から離れると急速に 0 に近づくようなものを選ぶ. このような ρ の典型例として

$$\rho(u) = \exp(-u/2) \quad (2)$$

などがあげられる.

この類似度を用いると **原則CP** は「各ノード i において, 隣接する (すなわち $a_{i,j} = 1$ となる) ノード j と自分との類似度 $\rho(d_{i,j})$ が, 隣接しない ($a_{i,k} = 0$ となる) ノード k との類似度 $\rho(d_{i,k})$ よりも大きくするように配置する」といい換えられる. すなわち,

$$\rho(d_{i,j}) > \rho(d_{i,k}) \quad (a_{i,j} = 1, a_{i,k} = 0 \text{ のとき}) \quad (3)$$

となる.

ここで $\rho(d_{i,j})$ を i, j の関数と見ると, 式 (3) は $\rho(d_{i,j})$ が $a_{i,j}$ の十分良い近似となっていればよい. この近似の度合いは, $a_{i,j}$ と $\rho(d_{i,j})$ のクロスエントロピー:

$$E_{i,j} = -a_{i,j} \ln \rho(d_{i,j}) - (1 - a_{i,j}) \ln(1 - \rho(d_{i,j})) \quad (4)$$

で評価できる. 実際式 (4) は $\rho(d_{i,j}) = a_{i,j}$ のとき, すなわち連続値類似度関数が離散類似度と完全に一致するとき最小となる. $E_{i,j}$ が i, j に関して対称であることに注意すると, $\mathbf{x}_1, \dots, \mathbf{x}_N$ に関し最小化すべき全エネルギー関数は次式となる.

$$E = \sum_{i=1}^{N-1} \sum_{j=i+1}^N E_{i,j}. \quad (5)$$

ここで, ノード i を固定すると, $a_{i,j}$ はノード j がノード i と同じクラスに属するか否かを示す 2 値クラ

スラベルと解釈することができる．すなわち $a_{i,j} = 1$ の場合， j は i と同じクラスに属し， $a_{i,j} = 0$ の場合 j は i とは異なるクラスに属すと見なす．したがってこの問題は，式 (5) を（分類問題としては標準的な）目的関数として， $\mathbf{x}_1, \dots, \mathbf{x}_N$ をパラメータとする N 個の 2 分類問題を同時に解く問題と見なせる．

式 (2) を用いると式 (4) は次式となる．

$$E_{i,j} = \frac{a_{i,j}}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - (1 - a_{i,j}) \ln(1 - \exp(-\frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2)) \quad (6)$$

最終的な目的関数としては，正則化のための weight-decay 項を付加した次式を用いる．

$$J = \sum_{i=1}^{N-1} \sum_{j=i+1}^N E_{i,j} + \frac{\mu}{2} \sum_{i=1}^N \|\mathbf{x}_i\|^2 \quad (7)$$

ここで μ は正則化係数であり，事前に決めておく．正則化項の導入によりアルゴリズムは安定し，かつ埋め込まれた配置全体のサイズが制御される．以降本論文では，この目的関数を用いる提案法を CoPE (Connectivity Preserving Embedding) 法と呼ぶ．

2.2 学習アルゴリズム

CoPE 法に基づく我々の学習アルゴリズムでは N 個の点すべてを同時に移動させるのではなく，KK バネモデルの学習アルゴリズムと同様に最も勾配ベクトルのノルムが大きい $\mathbf{x}_i$ だけを Newton-Raphson 法を用いて移動させる．このような学習方法は式 (7) のような正則化項付きのクロスエントロピーを用いるという点も含めて，分類問題を解くオンライン学習法としてニューラルネットワーク研究ではよく知られる手法である⁹⁾．

CoPE 学習アルゴリズムの概要を図 2 に示す．ここで $J_{\mathbf{x}_i}$ は式 (7) で定義される目的関数 J の $\mathbf{x}_i$ に関する勾配ベクトルであり，次式によって計算される．

$$J_{\mathbf{x}_i} = \frac{\partial J}{\partial \mathbf{x}_i} = \sum_{j \neq i} \frac{\partial E_{i,j}}{\partial \mathbf{x}_i} + \mu \mathbf{x}_i \quad (8)$$

$E_{i,j}$ の微分は次式によって計算できる．

$$\frac{\partial E_{i,j}}{\partial \mathbf{x}_i} = \frac{a_{i,j} - \exp(-\frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2)}{1 - \exp(-\frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2)} (\mathbf{x}_i - \mathbf{x}_j) \quad (9)$$

図中のステップ (2) において $t = 1$ における勾配ベクトルは N 個のすべての点について式 (8) に従って計算されるので，その計算量は $O(KN^2)$ である．一方，ステップ (6) での $J_{\mathbf{x}_i}$ の更新時において完全な再計算が必要となるのはステップ (3) で選ばれた $\mathbf{x}_i$ についてのみであり，他の $\mathbf{x}_j \neq \mathbf{x}_i$ については次式により差分のみ更新することで効率的に計算できる．

入力として $N \times N$ 隣接行列 $\mathbf{A} = (a_{i,j})$ ，埋め込みの次元 K ，終了条件を決める収束精度定数 ϵ を与える．

- (1) $t = 1$ とし， $\mathbf{x}_1, \dots, \mathbf{x}_N$ をランダムに初期化する．
- (2) 勾配ベクトル $J_{\mathbf{x}_1}^{(1)}, \dots, J_{\mathbf{x}_N}^{(1)}$ を計算する．
- (3) $\|J_{\mathbf{x}_i}^{(t)}\|$ が最大となる $\mathbf{x}_i$ を選択する．
- (4) もし $\|J_{\mathbf{x}_i}^{(t)}\| < \epsilon$ ならば $\mathbf{x}_1, \dots, \mathbf{x}_N$ を出力して終了する．
- (5) 変分ベクトル $\Delta \mathbf{x}_i$ を計算する．
- (6) $J_{\mathbf{x}_1}^{(t)}, \dots, J_{\mathbf{x}_N}^{(t)}$ を用いて $J_{\mathbf{x}_1}^{(t+1)}, \dots, J_{\mathbf{x}_N}^{(t+1)}$ を更新する．
- (7) $\mathbf{x}_i = \mathbf{x}_i + \Delta \mathbf{x}_i$ によって $\mathbf{x}_i$ を更新する．
- (8) $t = t + 1$ とし，ステップ (3) へ戻る．

図 2 CoPE 学習アルゴリズムの概要

Fig. 2 CoPE learning algorithm.

$$\begin{aligned} J_{\mathbf{x}_j}^{(t+1)} &= J_{\mathbf{x}_j}^{(t)} - \frac{\partial E_{i,j}^{(t)}}{\partial \mathbf{x}_j} + \frac{\partial E_{i,j}^{(t+1)}}{\partial \mathbf{x}_j} \\ &= J_{\mathbf{x}_j}^{(t)} + \frac{\partial E_{i,j}^{(t)}}{\partial \mathbf{x}_i} - \frac{\partial E_{i,j}^{(t+1)}}{\partial \mathbf{x}_i} \end{aligned} \quad (10)$$

ここで式 (10) の 1 行目および 2 行目の変形は次の関係式が成り立つことによる．

$$\frac{\partial E_{j,h}^{(t+1)}}{\partial \mathbf{x}_j} = \frac{\partial E_{j,h}^{(t)}}{\partial \mathbf{x}_j}, \quad \frac{\partial E_{j,i}^{(t)}}{\partial \mathbf{x}_j} = -\frac{\partial E_{i,j}^{(t+1)}}{\partial \mathbf{x}_i} \quad (11)$$

ただし $h \neq i, j$ とする．したがって，ステップ (6) は $O(KN)$ で計算できる．

ステップ (5) において，Newton-Raphson 法によれば $\Delta \mathbf{x}_i$ はヘッシアン $\mathbf{H}$ を用いて次式で計算される．

$$\Delta \mathbf{x}_i = -\mathbf{H}^{-1} J_{\mathbf{x}_i}^{(t)}, \quad \text{ただし} \quad \mathbf{H} = \frac{\partial^2 J^{(t)}}{\partial \mathbf{x}_i \partial \mathbf{x}_i'} \quad (12)$$

ここで $\mathbf{x}$ の転置行列を $\mathbf{x}'$ で表す．KK バネモデルにおいても式 (12) が用いられている．しかし，KK バネモデル，CoPE 法ともに， $\mathbf{H}$ は必ずしも正定値とは限らないため， $J^{(t+1)} < J^{(t)}$ がつねに成り立つとは限らない．したがって CoPE 法においては，式 (12) を適用した結果， J が増加した場合は式 (12) の適用をやめ，代わりに次の式 (13) を用いる．

$$\Delta \mathbf{x}_i = \lambda J_{\mathbf{x}_i}^{(t)} \quad (13)$$

ただしステップ長 λ は $J^{(t+1)} < J^{(t)}$ となるように選ぶ． $J_{\mathbf{x}_i}^{(t)}$ は勾配方向であるので，十分小さい λ を選ばばつねに J は減少する．この変更によって提案アルゴリズムの収束が保証される．なお J の更新は次式のように差分だけ計算することによって $O(KN)$ で計算できる．

$$\begin{aligned}\Delta J &= J^{(t+1)} - J^{(t)} \\ &= \sum_{j \neq i} (E_{i,j}^{(t+1)} - E_{i,j}^{(t)}) \\ &\quad + \frac{\mu}{2} \{ \|\mathbf{x}_i + \Delta \mathbf{x}_i\|^2 - \|\mathbf{x}_i\|^2 \}\end{aligned}\quad (14)$$

3. 実験による評価

3.1 従来の埋め込み法

実験を行うにあたって、提案する CoPE 法を Torgerson による古典的な多次元尺度法（古典 MDS）⁷⁾、スペクトラルクラスタリング⁸⁾そして KK パネモデルという代表的な 3 つの従来法と比較した。まずこれらについて簡単に説明する。

$\mathbf{G} = (g_{i,j})$ をグラフ距離を用いた任意の 2 ノード間の距離行列、 $\mathbf{O} = (o_{i,j})$ を $o_{i,j} = g_{i,j}^2$ で定義される行列、 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]'$ を埋め込みの結果得られた N 個の点の座標からなる $N \times K$ 行列とする。

古典 MDS では次式で与えられる目的関数の最小化を考える。

$$J_{\text{CMDS}} = \text{trace}\{(-\frac{1}{2}\mathbf{Y}\mathbf{O}\mathbf{Y} - \mathbf{X}\mathbf{X}')^2\} \quad (15)$$

ただし $\mathbf{Y}$ は N -次元 Young-Householder 変換行列である。

$\mathbf{B} = (b_{i,j})$ を $b_{i,j} = \exp(-g_{i,j}/2)$ ($i \neq j$ のとき)、 $b_{i,i} = 0$ で定義される類似度行列 (affinity matrix) とする。このときスペクトラルクラスタリングの目的関数は次式で定義される。

$$J_{\text{SC}} = \text{trace}\{(\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2} - \mathbf{X}\mathbf{X}')^2\} \quad (16)$$

ここで $\mathbf{D}$ は対角行列であって、 $\mathbf{D}$ の (i,i) 成分は $\mathbf{B}$ の第 i 行の全要素の和とする。この方法では $\mathbf{X}$ の各行の二乗和が 1 になるように次式によって正規化した球面上に結果を配置する。

$$\hat{x}_{i,j} = \frac{x_{i,j}}{\sqrt{\sum_{j=1}^N x_{i,j}^2}}. \quad (17)$$

一方 KK パネモデルでは、すでに述べたように任意のノード間にグラフ距離に対応する長さのばねを設定するため、目的関数は次式で定義される。

$$J_{\text{KK}} = \frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{(g_{i,j} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2}{g_{i,j}^2} \quad (18)$$

3.2 実験に用いたデータ

実データから構成した 3 種類の異なるタイプのネットワークを用いた評価実験を行った。生物学に由来する第 1 のネットワークは大腸菌 (*Escherichia coli*) の遺伝子制御ネットワーク¹⁰⁾であり、ここでは略して

表 1 各ネットワークの主な統計量
Table 1 Summary statistics of the networks.

NAME	N	L	$\bar{L}$	L°	$\bar{G}$	G°
E.Coli	328	456	2.78	72	4.83	13
NIPS	1061	2080	3.92	45	7.23	17
NTT	3870	9337	4.83	279	6.41	17

E.Coli と呼ぶ。第 2 のネットワークは国際会議 NIPS (Neural Information Processing Systems) の第 1~12 回に発表された論文の共著者ネットワーク¹¹⁾であり、略して **NIPS** と呼ぶ。**NIPS** は各著者をノードとし、2 人の著者が少なくとも 1 つの共著論文を発表すればその著者間にリンクを張ることで構成される。第 3 番目は WWW (World Wide Web) のハイパーリンクネットワークであり、NTT (日本電信電話株式会社) のサイト: www.ntt.co.jp ドメインに属する WWW ページをすべて収集して構成した。略して **NTT** と呼ぶ。

我々の実験ではこれらのネットワークをまず無向ネットワークに変換し、さらに最大連結成分のみを抜き出したものを用いた。通常このように各連結成分を別々に計算するのが自然である。表 1 に抜き出したネットワークに関する統計量を示す。ここでノード数を N とし、また、 L , $\bar{L}$, L° はそれぞれ各ネットワークの合計リンク数、平均リンク数、最大リンク数を表す。 L_i をノード i と直接つながっているリンク数とすると、これらの間には

$$\begin{aligned}L &= \frac{1}{2} \sum_{i=1}^N L_i, \quad \bar{L} = \frac{1}{N} \sum_{i=1}^N L_i, \\ L^\circ &= \max_i \{L_i\}\end{aligned}\quad (19)$$

という関係式が成り立つ。次に $\bar{G}$ および G° は各ネットワークにおける、任意の 2 ノード間のグラフ距離のそれぞれ平均値、最大値を表す。 $g_{i,j}$ をノード i と j の間のグラフ論的距離とすると次のような関係式が成り立つ。

$$\begin{aligned}\bar{G} &= \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N g_{i,j}, \\ G^\circ &= \max_{i,j} \{g_{i,j}\}\end{aligned}\quad (20)$$

これらのネットワークの隣接行列はすべて疎行列という点では共通しているが、各統計量についてはそれぞれ異なる特徴を持つことが分かる。また、**NTT** は **E.Coli**, **NIPS** に比べてサイズが特に大きく複雑なネットワークである。

3.3 評価尺度

今 N ノードからなる K 次元ユークリッド空間への埋め込みが得られたとし、その配置を $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ とする。そこで、各 $\mathbf{x}_i$ を中心とし、半径 r_i の K 次

元球体 $B_i(r_i)$ を考える。原則CP を完全に実現する理想的な埋め込みでは、各ノード i において適切な半径 r_i を選ぶことによって、 i に隣接する（直接リンクでつながっている）ノードをすべて含み、逆に隣接しないノードはまったく含まない、すなわちまったく違反がない $B_i(r_i)$ を構成できるはずである。しかしながら実際の埋め込みでは、特に次元 K が小さく複雑なネットワークの場合、すべてのノード i でそのような r_i を選ぶことは一般に不可能である。そこで各ノード i において適切な尺度のもとで最適な（すなわち違反の度合いが最も低い） r_i を選ぶことを考える。

特に疎なネットワークの場合、隣接するノードに比べて隣接しないノード数が圧倒的に多いため、 $B_i(r_i)$ に正しく含まれるノードの個数および $B_i(r_i)$ から正しく除外されるノードの個数の合計を考える通常の「精度」(accuracy) を尺度として用いるのは適切ではない。なぜなら、隣接するかどうかにかかわらずすべてのノードがまったく $B_i(r_i)$ に含まれない場合でも、精度は高くなるからである。そこでより適切な尺度として、情報検索やテキスト分類などの分野で広く用いられる F -尺度を用いる。

F -尺度は「適合率」(precision) と「再現率」(recall) との調和平均で定義される。今、 $\#X$ で集合 X の要素数を表すことにすると、 $\mathbf{x}_i$ と r_i に対応する i 番目の球体 $B_i(r_i)$ の適合率 $P_i(r_i)$ は、 $B_i(r_i)$ 内の $\mathbf{x}_i$ 以外のノードのうち $\mathbf{x}_i$ と同一クラスのノード数の比率、すなわち、

$$P_i(r_i) = \frac{\#\{j|\mathbf{x}_j \in B_i(r_i), a_{i,j} = 1, j \neq i\}}{\#\{j|\mathbf{x}_j \in B_i(r_i), j \neq i\}} \quad (21)$$

で定義される。再現率 $R_i(r_i)$ は $\mathbf{x}_i$ と同一クラスのノードが $B_i(r_i)$ 内に属する比率、すなわち、

$$R_i(r_i) = \frac{\#\{j|\mathbf{x}_j \in B_i(r_i), a_{i,j} = 1, j \neq i\}}{\#\{j|a_{i,j} = 1, j \neq i\}} \quad (22)$$

で定義される。適合率と再現率はトレードオフの関係にあり、適合率を高くするためには r_i を小さく、逆に再現率を高くするためには r_i を大きくする必要がある。したがって最適な半径 $\hat{r}_i$ は両者の調和平均で定義される以下の F -尺度を最大にする r_i として求まる。

$$F_i(r_i) = 1 / \left\{ \alpha \frac{1}{P_i(r_i)} + (1 - \alpha) \frac{1}{R_i(r_i)} \right\} \quad (23)$$

ただし α は重みで、本論文では $\alpha = 0.5$ とした。式 (23) をすべてのノード i について平均することによって、低次元に埋め込まれたネットワークの配置を定量的に評価する尺度である「接続 F -尺度」(connectivity

F -measure) を次式で定義する。

$$F = \sum_{i=1}^N \frac{F_i(\hat{r}_i)}{N} \quad (24)$$

我々はこの接続 F -尺度を主に CoPE 法による埋め込み結果の評価の基準として用いる。この評価基準は原則CP を特に考慮したものであるが、古典 MDS や KK パネモデルにおいて、仮に原則DP が完全に満足され、グラフ距離が低次元空間で完全に再現された場合は、各 i について $\hat{r}_i = 1$ とすることによって $F = 1$ が成り立つ。したがって、提案する評価基準は従来の埋め込み方法に対しても十分合理的な尺度であるといえる。

3.4 接続 F -尺度を用いた比較

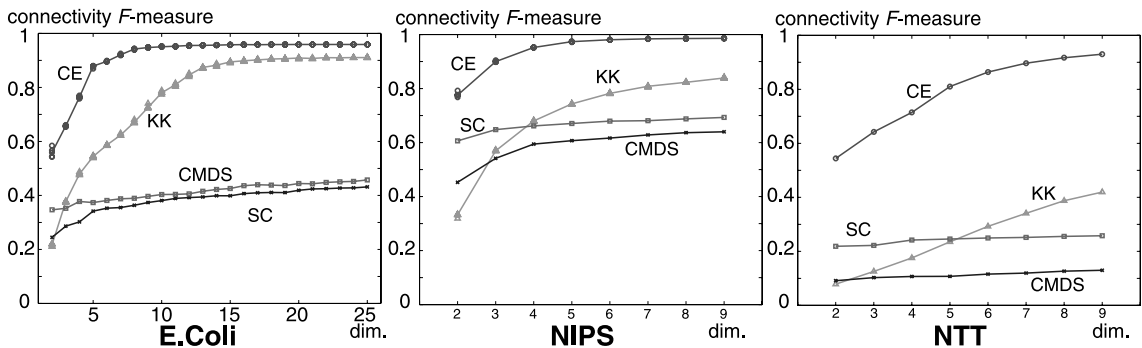
CoPE 法ならびに、古典 MDS、KK パネモデル、スペクトラルクラスタリングの 3 つの既存法を上述べた 3 種類のネットワークにそれぞれ適用し、 K 次元空間への埋め込みを行った。さらにその結果を接続 F -尺度を用いて評価した。 K の値について、E.Coli については 2 から 25 まで、NIPS と NTT については 2 から 9 までそれぞれ変化させて実験した。図 3 に評価結果のグラフを示す。縦軸は接続 F -尺度の値を、横軸は埋め込んだ次元 K を表す。CoPE 法および KK パネモデルについては、局所解の問題があるため乱数の初期値を変えて 5 回行った結果を表示しているが、結果を見る限り、結果のばらつきはほとんどなく、局所解の影響は無視できることが分かる。各手法とも、次元数 K および、埋め込みの自由度の増加とともに単調に接続 F -尺度が向上していることが分かる。

提案する CoPE 法が他の手法に比べて、特に配置が難しい低次元において、優れていることが分かる。なかでも最もサイズの大きい NTT では CoPE 法は従来法を顕著に上回っている。

3.5 2 次元への埋め込み結果

図 4 のそれぞれ左側に古典 MDS (CMDS)、KK パネモデル (KK)、CoPE 法 (CoPE) による NTT の実際の 2 次元への埋め込み結果を示す。スペクトラルクラスタリングは 3 次元球面上の配置となるため、比較を省略した。

すでに述べたように、これらの配置結果を「美観上好ましい」(aesthetically pleasing) かどうかによって評価するのは難しいが、特筆すべきいくつかの特徴的差異が観測できる。図 4 (a) に示す古典 MDS の結果は、多くのノードが 1 点に縮退してしまっているため、可視化、特にブラウジングの観点では問題がある。これは線型手法の限界と考えられる。一方、図 4 (b) に

図3 接続 F -尺度を用いた比較実験Fig. 3 Experimental comparisons of the connectivity F -measure.

示す KK バネモデル、および、図 4(c) に示す提案法である CoPE 法はこのノード縮退が少ない。さらにこれらを注意深く見ると、前者では多くのノードが半円状にかたまって配置されている（これはグラフ距離が離散値をとることに起因しており、「たんぼ効果」と呼ばれる⁴⁾）のに対し、後者ではより一様に放射状に広がっている。CoPE 法のほうが空間をより効率的に活用していることが分かり、この差が両者の接続 F -尺度の性能差として現れているともいえる。KK バネモデルによる配置結果は、特にノードが密集した領域において、ブラウジングを難しくしていると考ええる。

大規模で複雑なネットワークのブラウジングでは、特に注目したいネットワークの一部を、それ以外のノードおよびリンクを取り除くことによって切り出し、その部分に特化したより詳細なブラウジングを考える必要がある。その際再計算が必要となるが、それによって配置が大幅に変わってしまうのは望ましくない。本論文では、切り出しの前後で配置が大幅に変わらないことを切り出し安定性があるという。図 4 それぞれの左側の図において、点線で囲まれた領域を切り出し、この部分のみ再計算して埋め込んだのがそれぞれの右側の図である。切り出し前後を含む、合計 6 つの点線で囲まれた領域はすべて、同一のサブネットワークに対応している。容易に分かるように、古典 MDS、KK バネモデルにおいては、この切り出しの前後において、配置がかなり変化しているのに対し、CoPE 法では比較的变化が少なく、したがって切り出し安定性が高いと考えられる。以上の結果から、CoPE 法は従来法よりブラウジングに適した埋め込み方法であるといえる。

4. 関連研究

固有値解析に基づく方法は古典 MDS やスペクトラルクラスターリング以外にもいくつか存在する。たとえば Tenenbaum らによる Isomap¹²⁾ や Roweis ら

による LLE (Locally Linear Embedding)¹³⁾ などが知られている。しかし、これらの方法は、高次元ユークリッド空間に座標を持つ点の集合を低次元空間に次元圧縮して埋め込むことを目的としており、座標値が未知のリレーショナルデータを対象とする提案法とは本質的に異なる。最近 Hinton らによって提案された stochastic embedding¹⁴⁾ も高次元の座標値を前提としているという点で同様である。

一方、ばねモデルの考え方をういたグラフの埋め込みを提案したのは Eades が最初である³⁾。Eades のモデルでは、隣接ノード間は長さ 1、非隣接ノード間は長さ無限大のばねで拘束されていると考える。ただし、通常のフックの法則に従う線形のばねでは強すぎるので、対数のばねを用いるとしている。Eades の方法はその後のばねモデルの基礎となっており、すでに説明した KK バネモデルでは、任意のノード間がグラフ距離に等しい長さのばねで結ばれていると考える。なお現在では単に「ばねモデル」と呼ぶときは KK バネモデルを指すことがほとんどであり、Eades の方法はほとんど用いられない。たとえば著名なグラフ描画ソフトウェアのパッケージ graph-viz¹⁵⁾ の一部として配布されている NEATO というプログラムはそのコア部分に KK バネモデルが忠実に実装されている。

古典 MDS や KK バネモデルを含むほとんどのアルゴリズムは、任意のノード間に距離が定義されていることを前提としている。これは、たとえ疎 (sparse) なネットワークであっても、つねに疎でない巨大な距離行列を扱わなければならないことを意味しており、ノード数の増大にともなって使用メモリ量が爆発的に増加するという意味で、スケーラビリティ上問題である。一方、提案する CoPE 法においては隣接行列のみを扱う。隣接行列は疎なネットワークの場合、疎行列表現によりメモリの大幅な効率化が可能であるという点で有利である。

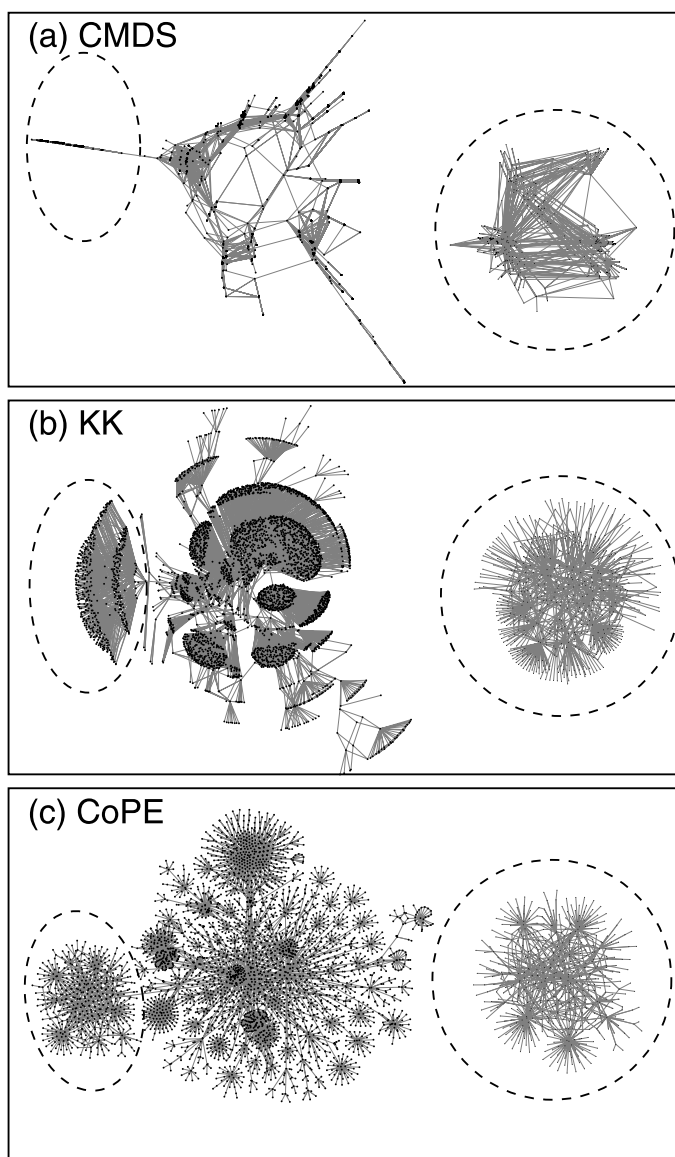


図 4 古典 MDS(a), KK パネモデル (b), CoPE 法 (c) による **NTT** の 2 次元への埋め込み結果. それぞれ左側が全体図, その中の点線部分を切り出して再計算したものを右側に示す

Fig. 4 2D layouts of **NTT** by the classical MDS (a), the KK spring method (b) and the CoPE method (C). Each picture in the left shows the whole network and the sub network in the dotted circle is re-embedded and shown in the right.

5. 結 論

本論文ではクロスエントロピー最小化に基づくネットワークデータの埋め込み法を新たに提案し, さらに, F -尺度に基づいて埋め込み結果を定量的に評価する新たな評価尺度を提案した. この尺度に基づき, 実際のネットワークデータを用いた実験を行った結果, 提案

法による埋め込み結果は従来法に比べて優れていることが分かった. また, 実際の埋め込み結果の比較においても, 我々の方法は空間をより効率的に利用し, また, 切り出し安定性にも優れており, よりブラウジングに適した埋め込みを実現できることを確認した. 今後は特に Web ブラウジングに焦点を絞り, Network Motif¹⁰⁾ などを用いた特徴分析も考慮した, WWW

向け発見システムへと発展させる予定である。

参 考 文 献

- 1) Ong, I.M., Glasner, J.D. and Page, D.: Modelling regulatory pathways in E.coli from time series expression profiles, *Journal of Bioinformatics*, Vol.18 Suppl., pp.S241-S248 (2002).
- 2) Wasserman, S. and Faust, K.: *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, UK (1994).
- 3) Eades, P.: A heuristic for graph drawing, *Congressus Numerantium*, Vol.42, pp.149-160 (1984).
- 4) Buja, A., Swayne, D.F., Littman, M., Dean, N. and Hofmann, H.: XGvis: Interactive Data Visualization with Multidimensional Scaling, Technical report, AT&T Labs-Research, Florham Park, NJ (2001).
- 5) Kamada, T. and Kawai, S.: An algorithm for drawing general undirected graph, *Information Processing Letters*, Vol.32, pp.7-15 (1989).
- 6) Floyd, R.W.: Algorithm 97: shortest path, *Comm. ACM*, Vol.5, p.345 (1962).
- 7) Torgerson, W.S.: *Theory and Methods of Scaling*, Wiley, New York (1958).
- 8) Ng, A., Jordan, M. and Weiss, Y.: On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems 14*, Vol.14, MIT Press, Cambridge, MA. (2002).
- 9) Bishop, C.M.: *Neural networks for pattern recognition*, Clarendon Press, Oxford (1995).
- 10) Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U.: Network motifs in the transcriptional regulation network of Escherichia coli, *Nature Genetics*, Vol.31, No.1, pp.64-68 (2002).
- 11) Roweis, S.T.: Data for MATLAB hackers: NIPS Conference Papers Vols 0-12 (2002). <http://www.cs.toronto.edu/~roweis/data.html>
- 12) Tenenbaum, J.B., de Silva, V. and Langford, J.C.: global geometric framework for nonlinear dimensionality reduction, *Science*, Vol.290, pp.2319-2323 (2000).
- 13) Roweis, S.T. and Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding, *Science*, Vol.290, pp.2323-2326 (2000).
- 14) Hinton, G. and Roweis, S.T.: Stochastic Neighbor Embedding, *Advances in Neural Information Processing Systems 15*, Vol.15, MIT

Press, Cambridge, MA. (in press).

- 15) North, S.C.: *NEATO User's Guide*, AT&A Bell Laboratories (1992).

(平成 15 年 4 月 7 日受付)

(平成 15 年 7 月 3 日採録)



山田 武士 (正会員)

昭和 39 年生。昭和 63 年 3 月東京大学理学部数学科卒業。同年 NTT 入社。平成 8 年より 1 年間英国コベントリー大学客員研究員。現在、NTT コミュニケーション科学研究所主任研究員。主として機械学習、メタヒューリスティクスによる組合せ最適化等の研究に従事。人工知能学会、電子情報通信学会各会員。



斉藤 和巳 (正会員)

昭和 38 年生。昭和 60 年慶応義塾大学理工学部数理科学科卒業。工学博士。同年 NTT 入社。平成 3 年より 1 年間オタワ大学客員研究員。神経回路網、機械学習の研究に従事。現在、NTT コミュニケーション科学基礎研究所主任研究員 (特別研究員)。情報処理学会論文賞受賞 (平成 9 年)。人工知能学会論文賞受賞 (平成 11 年)。電子情報通信学会、人工知能学会、日本神経回路学会、IEEE 各会員。



上田 修功 (正会員)

昭和 33 年生。昭和 57 年大阪大学工学部通信工学科卒業。昭和 59 年同大学院修士課程修了。同年 NTT 電気通信研究所入所。平成 5 年より 1 年間米国 Purdue 大学客員研究員。現在、NTT コミュニケーション科学基礎研究所知能情報研究部長 (創発学習研究グループリーダー兼務)。奈良先端科学技術大学院大学客員助教授。日本神経回路学会理事。画像処理、コンピュータビジョン、パターン認識、統計的学習、Web 統計解析の研究に従事。博士 (工学)。平成 4 年日本神経回路学会研究奨励賞、平成 9 年電気通信普及財団賞 (テレコムシステム技術賞)、平成 12 年電子情報通信学会論文賞受賞。電子情報通信学会、日本神経回路学会、IEEE 各会員。