

動的に競合層を構成する自己組織化マップの提案とその検証

日大生産工(院) ○檜山 昌弘 日大生産工 山内 ゆかり

1 まえがき

情報技術の急速な発達に伴い、世界中でデータの量が複雑かつ膨大になっていることから、データマイニングは重要な課題である。クラスタリングは代表的な分析手法の1つであり、自己組織化マップ (Self-Organizing Map: SOM)[1]の利用が注目されている。SOMは、1982年にKohonenによって提案された教師なしニューラルネットワークである。SOMはニューロン間のつながりである位相的順位を保持できるという利点を持ち、類似性のあるデータを分類するのに適している。SOMの改良は、ニューロンの役割による学習率の差別化[2]や、データの類似度を考慮した近傍関数[3]など、学習部分に関する研究が主流である。競合層の最適化に関する研究は、入力データから動的に競合層を構成する(Growing SOM: GSOM)[4]などの研究がある。これらの手法の評価には量子化誤差、トポロジー誤差、ニューロン利用率などの指標が用いられるが、各手法によりトレードオフがある。また、用いられる競合層は基本的に格子状である。

須藤らは、ニューラルネットワークの構造を動的に構成する自己増殖型ニューラルネットワーク (Self-Organizing Incremental Neural Network: SOINN)[5]を提唱し、予めデータの特徴やクラスのないクラスタリングでの有効性を示した。

本研究では、従来の格子状の競合層を用いるのではなく、須藤らの自己増殖の概念に基づき、SOMの学習と共に競合層を動的に構成する動的自己組織化マップを提案し、検証を行う。

2 従来研究

2.1 自己組織化マップ(SOM)はKohonenによって提案されたニューラルネットワークの一種で、入力層と競合層の二層構造で構成され、与えられた入力情報の類似度を競合層上のニューロン間の距離で表現するモデルである。基本となるSOMでは、競合層のニューロン(ノード)は二次元格子状に配置され、 M (データの次元数) 個の入力層のニューロン (ノード) と

の結合荷重 (参照ベクトル) を持つ。図1に二次元格子状の競合層の例を示す。

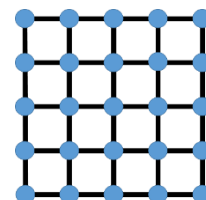


図1. 格子状の競合層

次にSOMのアルゴリズムを示す。

1. すべての競合層のノード i の参照ベクトル m_i をランダムな値で初期化する。
2. N 個の入力データからランダムに選んだものを入力ベクトル x とする。
3. 式(1)により入力ベクトル x と競合層のノード i の参照ベクトル m_i とのユークリッド距離を求め、その距離が最小となる勝者ノード c を見つける。

$$|x - m_c| = \min |x - m_i| \quad (i = 1, 2, \dots, M) \quad (1)$$

4. 式(2)を用いて、勝者ノードとその近傍の参照ベクトル m_i を更新する。ここで h_{ci} は近傍関数と呼ばれ、式(3)で定義される。

$$m_i(t+1) = m_i(t) + h_{ci}(t)m_i[x(t) - m_i(t)] \quad (2)$$

$$h_{ci}(t) = \alpha(t) \exp\left(-\frac{|r_c - r_i|^2}{2\sigma^2(t)}\right) \quad (3)$$

2.2 自己組織化マップの評価指標

SOMの学習及びデータ構造の表現能力を定量的に比較するために、量子化誤差、トポロジー誤差、ニューロン利用率の3つの指標を用いる。次にそれぞれの指標の説明を行う。

・量子化誤差 Qe

入力ベクトルとその勝者との距離の平均を計算した値である。 Qe の導出式を式(4)に示す。

$$Qe = \frac{1}{N} \sum_{j=1}^N \|x_j - m_c\| \quad (4)$$

ここで、 N は入力データの個数、 $\|x_j - m_c\|$ は入力データ x_j と勝者ニューロンの参照ベクトル

A Proposal and Validation of Dynamic Configuration Method of Competitive Layer for the Self-Organizing Map

Masahiro HIYAMA and Yukari YAMAUCHI

ル m_c のユークリッド距離である。従って、 Q_e は0に近いほど、入力データの特徴が競合層で表現されていることを示す。

・トポロジー誤差 Te

SOMの競合層が入力データの位相を表現できているかを示す値である。 Te の導出式を式(5)に示す。

$$Te = \frac{1}{N} \sum_{j=1}^N v(x_j) \quad (5)$$

入力データ x_j に対する1番目の勝者と2番目の勝者が互いに1近傍以内なら $v(x_j)$ は0、それ以外なら1となる。つまり Te は0に近いほど、類似したノードが近傍に配置され、競合層がデータ空間においてねじれた構造となっていないことを示す。

・ニューロン利用率 U

競合層のノードが勝者として選ばれる、活性ニューロンである割合を示す値である。 U の導出式を式(6)に示す。

$$U = \frac{1}{L} \sum_{i=1}^L u(i) \quad (6)$$

ここで、 L は競合層のノード数である。 $u(i)$ は、ノード i が1度以上勝者として選ばれた場合1、そうでない場合0となる。つまり、 U は1に近いほどより多くのニューロンを有効利用できており、不活性ニューロンが少ないことを示す。

2.3 類似度を考慮した自己組織化マップ

データの類似度を考慮した近傍関数を用いる研究に、竹内らが提案した自己組織化マップ(TSOM)がある。従来の近傍の概念は競合層上の距離に基づいているが、TSOMでは参照ベクトルの距離も考慮して近傍関数を定義する。TSOMの概念を図3に、近傍関数を式(7)に示す。

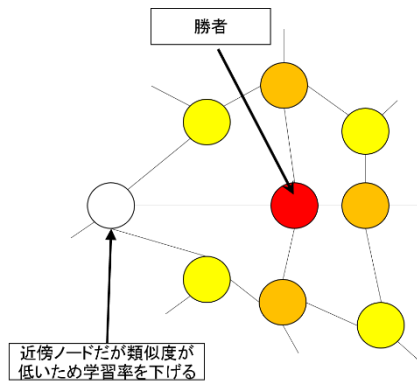


図2. TSOMの概念

$$h_{ci}(t) = \alpha(t) \exp\left(-\frac{|r_c - r_i|^2}{2\sigma^2(t)}\right) \times \exp\left(-\frac{|x_c - w_i|^2}{2\sigma^2(t)}\right) \quad (7)$$

$|x_c - w_i|^2$ は入力ベクトルと参照ベクトルの特徴差を表している。ただし、勝者ノードの学習は従来の近傍関数による式(2)を用いる。

TSOMは、従来のSOMと比べ、近傍範囲内であっても、学習の後期では類似度が低いと学習率が低くなる。これはその競合層上のノードの持つ特徴（個性）が近傍学習により失われることを防止し、局所的なデータの構造を維持する効果がある。そのため不活性ニューロンの割合が減り、ニューロン利用率 U の向上に繋がる。

図3(b)に示すAtomデータを用いたSOMとTSOMの3つの指標による比較を表1に示す。

表 1. Atom データの結果

	SOM	TSOM
Q_e	0.050	0.049
Te	0.214	0.235
U	0.865	0.937

類似度を考慮したため、不活性ニューロンが減少し、ニューロン利用率 U の向上に繋がることを確認できた。しかしトポロジー誤差 Te が悪化する結果となっている。

2.4 自己増殖型ニューラルネットワーク

須藤らに提唱された自己増殖型ニューラルネットワーク (Self-Organizing Incremental Neural Network: SOINN)は、入力された情報を元にノードの追加や削除、エッジの追加や削除を繰り返し行うことによって、情報数や分類数が不明な状況でも適切なクラスタリングを行う。SOINNのアルゴリズムを下記に示す。

1. 入力データをランダムに2つ選択し、その値を参照ベクトル m_j 、とする競合層のノードを生成する。
2. ランダムに選んだ入力データを入力ベクトル x とし、式(1)により入力データとの距離が最小となるノード c_1 を第1勝者、2番目に小さいノード c_2 を第2勝者とする。
3. 2つの勝者ノードの閾値 d_{c_1}, d_{c_2} をそれぞれ式(8)により算出する。

$$d_i = \begin{cases} \max_{k\text{-th node} \in N_i} \|m_i - m_k\| & (N_i \neq \emptyset) \\ \min_{k\text{-th node} \in A} \|m_i - m_k\| & (N_i = \emptyset) \end{cases} \quad (8)$$

ここで、 N_i はノード i とエッジで結合されているノードの集合、 A は競合層上の全てのノードの集合である。

4. 式(9)のどちらかが成立しない場合は入力ベクトルをノードとして競合層に追加する。どちらも成り立つ場合はノードを生成せず、第1、第2勝者をエッジで結合する。ただし元々エッジが引かれていた場合は、エッジの年齢を更新する。

$$\|x - m_{c_1}\| > d_{c_1}, \|x - m_{c_2}\| > d_{c_2} \quad (9)$$

5. ノード*i*が勝者に選ばれた回数 μ_i に基づき、第1勝者は式(10)、第1勝者の隣接ノードは式(11)を用いて参照ベクトルを更新する。

$$\Delta m_{c_1} = \frac{1}{\mu_{c_1}} (x - m_{c_1}) \quad (10)$$

$$\Delta m_i = \frac{1}{100\mu_i} (x - m_i) \quad (11)$$

6. 2-5を学習回数分繰り返す、ネットワークの更新周期 λ 毎に、少ないエッジを持つノードと勝者ノード間に選ばれないエッジの削除を行う。

SOINNでは、あらかじめネットワークの構造を決定する必要がなく、入力されたデータをノードとして競合層を構築することで、データの特徴を捉えたクラスタリングを行うことが可能である。また、従来のSOMと比較して、近傍学習の範囲が狭い（第1勝者の1近傍ノード以外は学習を行わない）ため、ノードの生成時の参照ベクトルの値が維持され、局所的構造の表現に適している。これはTSOMと同様に、高いニューロン利用率をもたらす。

しかし、ネットワーク構造はノードやエッジを削除する周期などのパラメータに依存し、適切にチューニングを行わないと汎化学習の性能が低下する。表2に図3(b)に示すAtomデータを用いたSOMとSOINNとの3つの指標による比較結果を示す。

表 2. Atom データの結果

	SOM	SOINN
Qe	0.050	0.100
Te	0.214	0.010
U	0.865	0.892

表2より、SOMと比較して、トポロジー誤差Teの大幅な改善、及びニューロン利用率Uの向上が確認できる。これは競合層を動的に構成することで、ランダムな初期配置や、データ空間と競合層の次元の違いによる、データ空間での競合層のねじれを改善しているからだと考えられる。しかし、競合層のノードの削除により局所的な特徴が失われ、量子化誤差Qeが悪化する結果となった。

3 提案手法

本研究では、SOINNの自己増殖のアルゴリズムとTSOMの類似度を考慮した学習を組み合わせ、動的に競合層を構成し、データの局所構造の特徴を維持する、動的自己組織化マップ(Dynamic TSOM: DTSOM)を提案する。下記に提案手法のアルゴリズムを示す。

1. 入力データをランダムに2つ選択し、その値を参照ベクトル m_j 、とする競合層のノードを生成する。
2. ランダムに選んだ入力データを入力ベクトル x とし、式(1)により入力データとの距離が最小となるノード c_1 を第1勝者、2番目に小さいノード c_2 を第2勝者とする。
3. 2つの勝者ノードの閾値 d_{c_1}, d_{c_2} をそれぞれ式(8)により算出する。
4. 式(9)のどちらかが成立しない場合は入力ベクトルをノードとして競合層に追加し、新しく生成されたノードと第1、第2勝者間をそれぞれエッジで結合する。どちらも成り立つ場合はノードを生成せず、第1、第2勝者をエッジで結合する。ただし、1つのノードがもつエッジの最大数は6とした。
5. 式(2)を用いて、勝者ノードとその近傍の参照ベクトル m_i を更新する。ただし、近傍関数は式(7)を用いる。ここで、 $\sigma(t)$ はネットワークの最長経路長に基づき決定する。
6. 2-5を学習回数分繰り返す。

4 実験環境

本実験では、図3に示す(a)Targetデータ、(b)Atomデータ、(c)Chainlinkデータの3つのデータ[6]を用いて、実験を行った。各データのデータ数、次元数、クラス数を表3、実験環境を表4に示す。本実験の実験結果は、それぞれのデータごとにシミュレーションを100回行い、その平均を表している。実験結果を表5、6、7に示す。

表 3. 実験に用いたデータ

	Target	Atom	Chainlink
次元数	2	3	3
データ数	770	800	1000
クラス数	6	2	2

表 4. 実験の環境設定

学習回数	データ*20
ノード数	100
エッジの上限年齢	データ数/4
ノード削除周期	データ数
学習率	0.5

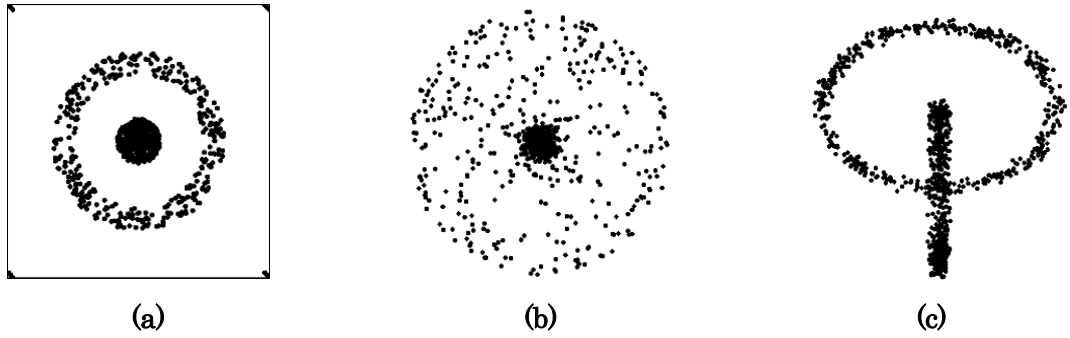


図3. (a)Targetデータ (b)Atomデータ (c)Chainlinkデータ

表 5. Target データの結果

	SOM	TSOM	SOINN	DTSOM
Qe	0.018	0.018	0.035	0.025
Te	0.109	0.141	0.016	0.015
U	0.808	0.857	1.000	0.965

表 6. Atom データの結果

	SOM	TSOM	SOINN	DTSOM
Qe	0.050	0.049	0.100	0.091
Te	0.215	0.235	0.010	0.040
U	0.864	0.934	0.892	0.994

表 7. Chainlink データの結果

	SOM	TSOM	SOINN	DTSOM
Qe	0.029	0.027	0.064	0.049
Te	0.267	0.310	0.016	0.013
U	0.769	0.895	0.864	0.993

表5から7より、提案手法であるDTSOMは、SOM及びTSOMと比較して、 Te の改善、及び U の向上が確認できた。これは動的に競合層を構成することで、データ空間におけるねじれを改善し、また類似度を考慮した学習により、局所データの特徴を維持したことで、不活性ニューロンの減少に繋がっていると考えられる。しかし、 Qe は悪化するという結果となった。またSOINNとの比較においては、 Qe の改善が確認できた。

以上により提案手法DTSOMは、SOINNの特徴である、トポロジー誤差 Te の大幅な改善、及びTSOMの特徴であるニューロン利用率 U の向上を実現する手法であることが確認できた。

5 まとめ

本研究では、TSOMの特徴である不活性ニューロンの減少とSOINNのトポロジー誤差が少ないという特徴を反映した、動的自己組織化マップの提案し、その有効性を検証した。

提案手法DTSOMにおいては、トポロジー誤差 Te の大幅な改善、及びTSOMの特徴であるニューロン利用率 U の向上を実現することができた。このことから、様々なデータに対し、従来の二次元格子状の競合層でクラスタリングを行うよりも、学習と共に競合層を構成することが有効であることが示された。今後の課題として、よりデータの特徴を反映したクラスタリングを行う為の手法の改善と、高次元かつ大規模な実データでの実験を行い、本研究の更なる有効性を確認することが挙げられる。

「参考文献」

- 1) T.Kohonen、 Self-Organizing Maps、Springer、 vol.30、 (1995)
- 2) 原口卓、 松下春奈、 西尾芳文:「怠け者を考慮したLazy Self-Organizing Mapとその振る舞い」、 電子情報通信学会技術研究報告. NLP、 非線形問題 108(174)、 5-10、 2008
- 3) 竹内良、 山内ゆかり:「自己組織化マップと交差点特徴を用いた手書き文字認識」日本大学生産工学部第46回学術講演会P-13 (2013)
- 4) Daminda Alahakoon、 Saman K. Halgamuge、 Bala Srinivasan: Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery、 IEEE Trans Neural Networks、 11(3)、 601-614 、 (2000)
- 5) 須藤明人、 佐藤彰祥、 長谷川修:「自己増殖型ニューラルネットワークを用いたノイズのある環境下での追加学習が可能な連想記憶システム」日本神経回路学会誌 Vol.15、 No.2、 98-109、 (2008)
- 6) Data Philipps Universitat Marburg Datenbionik (AGUltsch): http://www.uni-marburg.de/fb12/datenbionik/data?language_sync=1, (最終アクセス日時: 2015年10月21日 12時0分)