Vol.2010-MPS-77 No.22
2010/3/5

サンプルの所属度に応じた可変自己組織化マップ

多賀谷 侑史[†1] 安藤 晋[†1] 関庸一[†1]

標準の SOM では、2 次元平面上の指定した形状のマップ上にサンプルを配分する。その場合、サンプルが配分される利用ノードの配置があらかじめ決まっているので、サンプルが持つ自然な位相構造はマップの形状に折り畳まれ、少数のサンプルしか分類されない無駄なノードが生じることがある。それに対して本論文では、サンプルがノードに属している程度を表す所属度を定義し、それに応じて利用するノードの配置を変更する SOM の改良を提案する。利用ノードの数を変えずにその配置のみを変更することにより、サンプルが持つ位相構造をマップ上に自然に表すことができ、限られたノード数で不適合度を改善することができる。

Flexible Self-Organizing Maps adapted to degree of membership

Yuji Tagaya,[†1] Shin Ando[†1] and Yoichi Seki[†1]

Standard Self-Organizing Maps(SOM) distribute samples on fixed configuration of nodes. Then, a intrinsic topological structure of samples is folded, and some nodes have few samples. The nodes are useless. We propose a improvement of SOM. The improved SOM move its nodes adapted to degree of membership of samples. It can present intrinsic topological structure of samples and improve a degree of fitness.

1. はじめに

本研究では、SOM(Self-Organizing Maps, 自己組織化マップ)[1] を拡張して、高次元特徴量データのもつ自然な位相構造を、低次元の格子空間の中に表現する可変自己組織化マップを提案する。

高次元の特徴量を低次元に次元縮約する古典的な方法としては、主成分分析 (PCA)[2] や多次元尺度構成法 (MDS)[3] などがある。特徴量空間の直交変換を行う PCA や、サンプル間距離を再現する低次元空間を構成する古典的 MDS では、全サンプルを一括して扱うため、サンプルセットの大域的関係も保存した次元縮約が行われる。これに対して、LLE[4] や Isomap[5] は、特徴量空間の近傍ごとにその構造を抽出することにより、積極的に大域的構造を捨て、本質的な次元数の空間に縮約することを目指している。このうち、LLE では、各サンプルを近傍内で線形近似する重みを元に線形空間を構成する。また、Isomap は、近傍内での距離のみから測地線に沿った距離関係を求め、これから次元縮約された空間を構成する。これらの方法は、局所的な位相空間を接続することで、データに本質的な次元数の多様体を構成する方法を提供している。

SOM も同様に大域的距離関係を無視し、局所的な構造を接続することで次元縮約する方法である。通常は、有界な二次元格子を用意し、これを次元縮約した空間として用いる。その各格子点 (ノード) に特徴量空間の代表点 (参照ベクトル) を対応させ、各サンプルが、特徴量空間内で最近隣である代表点 (参照ベクトル) を持つ格子点 (最整合ノード) に対応付けられることで、格子空間への次元縮約が実現される。参照ベクトルは、格子空間での近傍格子点に所属するサンプルの重みづけ平均となるように、後述する収束算法で与えられる。得られた格子空間を SOM マップと呼ぶ。

SOM は、簡潔な算法でもって、高次元特徴量をもつサンプルセットを文節化し、二次元マップとして可視化する方法となる。また、各種の応用があり、例えば、時系列変化を可視化するため、時間的近傍を導入する方法[6] もある。LLE や Isomap と異なり、結果が、格子点上に文節化されたサンプルセットとなるため、離散的な取扱いが可能となり、多群の特徴量の関係をモデル化する基盤をあたえる方法としても、応用上有用な結果を与えている[7)–9)]。

しかし SOM では、サンプルを配分する二次元格子の領域形状を事前に指定することが必要となる。その指定された領域形状の中に全サンプルが対応づけられるため、データの持つ自然な位相構造がその形状に合うように折り畳まれ、詰め込まれることになる。この場合、大きく異なる参照ベクトルが隣接して配置され、さらに、両者の中間の特徴をもったサンプルが少ない場合には、少数のサンプルのみが配分されるノードが生ずることになる。また、本来、異なるクラスターに属すサンプル群が隣接することも生ずる。このようなクラスターを識別する手法として、画像認識の分野でラベリングを利用する方法[10),11)] などが提案されているが、複雑なパラメータ設定を必要とするという課題が残されている。さらに、多群の特徴量の関係をモデル化する基盤として SOM を利用する場合には、SOM マップ上で隣接

†1 群馬大学工学研究科情報工学専攻

するノードが類似した参照ベクトルを持つような次元縮約が望まれるが、自然な位相構造が折り畳まれて詰め込まれる部分では、この性質が成立しないこととなる。

本研究では、以上のような SOM の課題を解決するため、無限に広がる格子空間中で、任意形状の格子点集合の利用を許容する SOM の算法を提案する。この際、利用格子点の個数は、データの多様性を考慮して事前に指定するものとし、格子空間中のどのノードを利用するかについては、サンプルのノードへの所属度という提案指標に基づき、所属度累計の大きいノードを利用するという方法をあたえる。以上の方法により、限られた利用ノード数で不適合度を改善し、サンプルセットのもつ位相構造に合ったマップを作成できるように、SOM の拡張を行う。

以下では、まず、第 2 節で標準の SOM 算法とその問題点を数値例から示す。第 3 節では提案する SOM の算法を示し、その数値例を第 4 節で示す。第 5 節で収束パラメータの選択のための数値実験結果を示し、第 6 節で実データでの解析例としてクレジットカード利用履歴を分析した事例を示す。

2. 標準の SOM 算法とその問題点

Kohonen が提案した自己組織化マップでは、格子として有界な四角格子または六角格子を通常用い、各格子点をノードと呼ぶ。次元縮約された格子空間の格子点座標を $\mathbf{r}_k$ $(k = 1, 2, \ldots, K)$ とし、この空間上のノルムを $||\mathbf{r}||$ と表す。また、p 次元特徴量空間中のサンプルを $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^t$ $(i = 1, \ldots, N)$ とし、この空間上の距離を $d(\mathbf{x}, \mathbf{x}')$ と表すものとする。さらに、この空間中のノード k と対応づけられる代表点を表す参照ベクトルを $\mathbf{m}_k = (m_{k1}, m_{k2}, \ldots, m_{kp})^t$ とする。

2.1 SOM の算法

SOM は図 1 の算法でサンプルのノードへの配分を行う。本研究では、最整合ノード c とノード k のマップ上での近さに関する学習率関数 $h_{ck}(t)$ として、次式のガウス関数を用いる。

$$h_{ck}(t) = \alpha^{(t)} \cdot \exp(-||\mathbf{r}_c - \mathbf{r}_k||^2 / 2\sigma^{(t)2}) \tag{1}$$

ただし、スカラー $\sigma^{(t)} > 0$ と $\alpha^{(t)} > 0$ は t の単調減少関数であり、近傍半径、学習率係数と呼ぶ。本研究では、これらを次式の単調減少等差数列とする。

$$\sigma^{(t)} = \sigma_0 + (\sigma^{(0)} - \sigma_0)\,((T-t)/(T)) \tag{2}$$

$$\alpha^{(t)} = \alpha^{(0)}\,((T-t)/(T)) \tag{3}$$

$SOM(\mathbf{X},\ \mathbf{L},\ \{\mathbf{m}_k^{(0)}\},\ T,\ h_{ck}())$

```
1   for t = 0 to T
2     i = 一様乱数 on {1, ..., N};
3     x^(t) = x_i;
4     c = arg min_{k∈L} d(x^(t), m_k^(t));
5     m_k^(t+1) = m_k^(t) + h_ck(t)(x^(t) − m_k^(t));
7   return {m_k^(T)};
```

- $\mathbf{X}$; 特徴量データ ($\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n]^t$)
- $\mathbf{L}$; 最整合ノード選択対象として利用する格子点集合
- $\{\mathbf{m}_k^{(0)}\}$; 参照ベクトル初期値は通常乱数で与える。
- T ; 反復回数
- $h_{ck}(t)$; 最整合ノードが c である場合のノード k での学習率関数 $(0 < h_{ck}(t) < 1)$, t の単調減少関数。ノード c の近傍外では、0 とすることもある。

図 1 SOM アルゴリズム

ここで、σ_0 は十分小さな正の実数である。これにより近傍関数の指数の発散を避けている。

以上により、学習率関数 $h_{ck}(t)$ を $\sigma^{(0)}$、σ_0、$\alpha^{(0)}$、T から定めていることとなる。以下では、標準の SOM の呼出しを次で表す。$SOM(\mathbf{X},\ L,\ \{\mathbf{m}_k^{(0)}\},\ T,\ \{\alpha^{(0)}, \sigma^{(0)}, \sigma_0\})$

2.2 標準の SOM によるマップの問題点

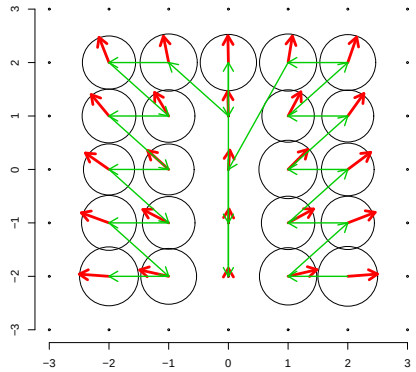
1 次元の位相構造を持つサンプルセットに SOM を適用した結果得られたマップを図 2 に示す。$[0, \pi]$ 上の一様乱数に従う θ から得られた $(\sin\theta, \cos\theta)^t$ をサンプル特性量としたデータ 10000 件をサンプルセットの数値例としている。2 次元の特徴量を持つ入力データであるが、それらを決定しているパラメータは θ で、1 次元であることから、入力データは 1 次元の位相構造を持つこととなる。マップは 5×5 の正方格子を用いている。図 2 の中央付近には、サンプルがほとんど所属しておらずデータの代表点として機能しない無駄なノードが存在している。本研究では、このような機能していないノードが自然な位相構造を折り畳んで詰め込んだことにより生じたものと考え、領域の形状の変形を許すことによりこのようなノードの生成を解消できる算法を提案する。

3. 可変自己組織化マップ

3.1 マップの拡張

無限に広がる正方格子または六角格子等の格子空間をマップとして用いる。この上でのノード集合に関する概念を図 3 に示す。各ノード k に対し、格子形状から定まる近傍を $\mathbf{N}_k$ とする。提案手法では、無限格子空間中から指定する個数 K 個のノードを選び、反復して

Vol.2010-MPS-77 No.22
2010/3/5



太い矢印は参照ベクトルであり、細い矢印は θ の値が大きく、かつ最も近いノードへ向けて引かれ位相構造を表している。円の大きさは、ノードに分類されたサンプル数を表している。$T = 10^5, \alpha^{(0)} = 0.02, \sigma^{(0)} = 1, \sigma_0 = 0.01, K = 25$。

図 2 通常 SOM で得られたマップ

更新する。第 s 回目に選択されたノード集合を $\mathbf{L}_s$ とする。利用ノード集合の初期値 $\mathbf{L}_1$ は適当な連結格子点集合とし、以下の算法により利用ノード集合を更新する。利用ノード集合の近傍を $\mathbf{U}_s = \cup_{k \in \mathbf{L}_s} \mathbf{N}_k$ と定め、利用候補ノード集合とする。

3.2 可変自己組織化マップアルゴリズム

可変自己組織化マップの算法を図 4 に示す。サンプル i がノード k に属している程度を (4) 式で定義し、所属度 b_{ik} とする。ただし、最整合ノードは利用ノード集合から選ぶこととする ($c_i = \arg\min_{k \in \mathbf{L}_s} d(\mathbf{x}, \mathbf{m})$)。

$$b_{ik} = \begin{cases} \frac{d(\mathbf{x}_i, \mathbf{m}_k)^{-2}}{\sum_{h \in \mathbf{N}_{c_i}} d(\mathbf{x}_i, \mathbf{m}_h)^{-2}} : k \in \mathbf{N}_{c_i} \\ 0 : k \notin \mathbf{N}_{c_i} \end{cases} \tag{4}$$

また、ノード k に割り振られた所属度の累計を $b_k = \sum_{i=1}^{N} b_{ik}$ とする。

4. 数 値 実 験

得られたマップがサンプルセットに合っている程度の評価には式 (5) を用い、これを不適合度と呼ぶ。また、本論文では正方格子を用い、ノード k に対してマップ上でユークリッド

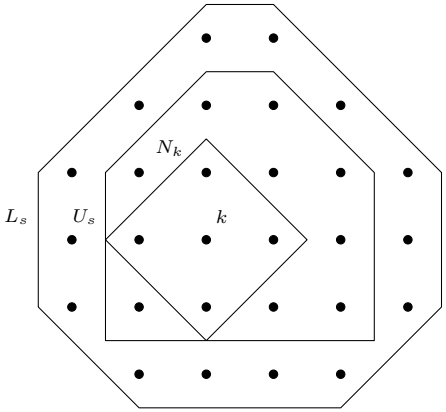


図 3 ノード集合の概念

$FlexibleSOM(\mathbf{X}, \mathbf{L}_0, \{\mathbf{m}_k^{(0)}\}, T, S, \{\alpha^{(0)}, \sigma^{(0)}, \sigma_0\})$

1 for $s = 0$ to S
2 $\quad \mathbf{U}_s = \cup_{k \in \mathbf{L}_s} \mathbf{N}_k$
3 $\quad \alpha_s^{(0)} = \alpha^{(0)} \left(\frac{S-s}{S}\right)$
4 $\quad \sigma_s^{(0)} = \sigma_0 + (\sigma^{(0)} - \sigma_0) \left(\frac{S-s}{S}\right)$
5 $\quad \{\mathbf{m}_k^{(0)}\} = SOM(\mathbf{X}, \mathbf{L}_s, T, \{\alpha_s^{(0)}, \sigma_s^{(0)}, \alpha_0\})$
6 $\quad \{b_k \mid k \in \mathbf{U}_s\}$ の算出
7 $\quad \mathbf{L}_{s+1} = \{k \in \mathbf{U}_s \mid b_k$ が上位 K 個 $\}$
8 return $(\mathbf{L}_S, \{\mathbf{m}_k^{(0)}\})$;

- L_0 ; 利用する初期格子点集合
- $\{\mathbf{m}_k^{(0)}\}$; 参照ベクトル ($k \in \mathbf{U}_0$)。通常、初期値は乱数値とする。
- T ; 標準 SOM 反復回数
- S ; 利用格子点の更新回数
- $\mathbf{X}$、$\{\alpha^{(0)}, \sigma^{(0)}, \sigma_0\}$; 2.1 節参照

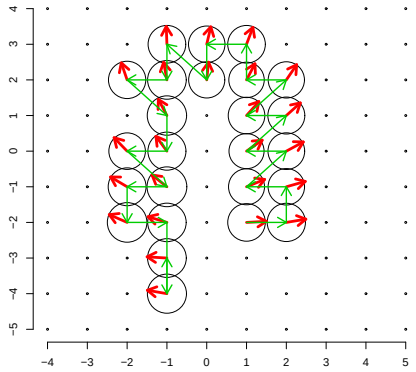
図 4 可変自己組織化マップアルゴリズム

距離が 1 以下のノード集合を近傍 $\mathbf{N}_k$ とする。

$$D^2 = \sum_{i=1}^{N} \min_k ||\mathbf{x}_i - \mathbf{m}_k||^2 \tag{5}$$

4.1 1 次元位相構造数値例の結果比較

2 節で使用した数値例に提案法を適用した結果を図 5 に示す。これから、使用している全てのノードにサンプルデータがほとんど均等に所属していることが分かる。 次に、不適合



記号は図 2 に同じ。$\alpha^{(0)} = 0.02, \sigma^{(0)} = 1, \sigma_0 = 0.01, T = 10^5, S = 100, K = 25$

図 5 可変自己組織化マップで得られたマップ

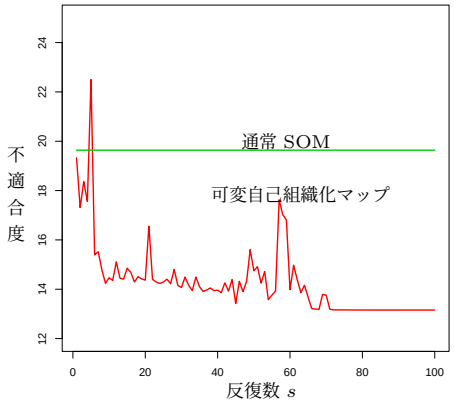


図 6 提案手法と通常の SOM の不適合度比較

度を通常の SOM と比較した結果を図 6 に示す。可変自己組織化マップの方が、反復数が増えるに従い不適合度が改善されている。また、反復の際に移動したノード数を図 7 に表す。反復数が 70 付近以降でノードの移動数は 0 となり、マップの形状が収束している。

4.2 2 次元位相構造数値例の結果比較

2 次元位相構造を持ち、複数の群が混合したサンプルセットから、その群を識別したマップが得られることを確認する。サンプルセットは、中心が標準偏差の 3 倍離れ、識別が容易な次の 3 つの分布から 10000 個ずつを抽出した混合サンプルセットを用意した。

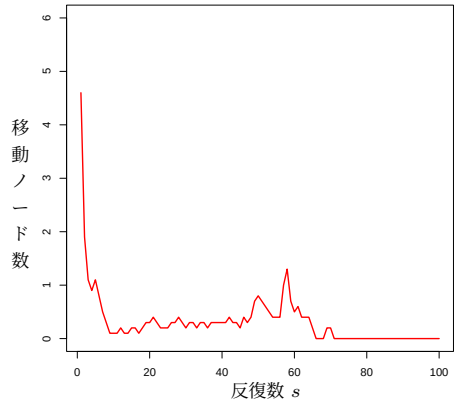
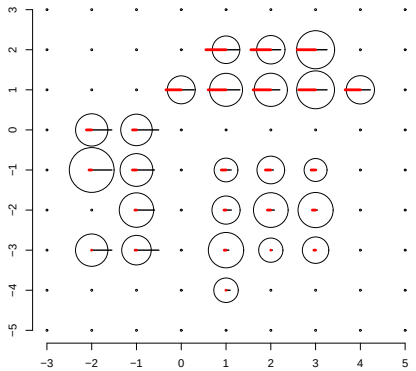


図 7 可変自己組織化マップ反復の際の移動ノード数



細い線が x、太い線が y。

図 8 2 次元サンプルセットから得られたマップ

$$Group1: \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{I} \right) \tag{6}$$

$$Group2: \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 3 \\ 0 \end{pmatrix}, \mathbf{I} \right) \tag{7}$$

$$Group3: \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 1.5 \\ 1.5\sqrt{3} \end{pmatrix}, \mathbf{I} \right) \tag{8}$$

得られたマップ (図 8) では、ノード集合の近傍が他のノード集合と重ならない 3 つのクラ

スターに分かれている。各ノードクラスターに属するサンプル数を調べると、それぞれの群のサンプルは完全に分離できていることが分かる。つまり、予備知識なしで、3 つの分布を、近傍関係で連結である 3 つのクラスターとして分類できたことが分かる。

5. パラメータの選択

本節では、収束パラメータの設定によって得られる不適合度の相違を 2 節の数値例で調べる。

5.1 近傍半径と学習率係数

近傍関数の近傍半径と学習率係数の 2 つのパラメータをそれぞれ 3 段階で変化させて提案法を行い、結果を比較する。近傍半径 $\sigma_0^{(0)}$ は、$1, 3, 5$ の 3 段階とする。最小の隣接ノード間の距離は 1 であるので、それを最小値とした。また、25 個のノードを正方形に配置した場合、一度に大半のノードを含む距離である 5 を最大値とした。学習率係数 $\alpha_0^{(0)}$ は、$0.02, 0.5, 0.9$ の 3 段階とした。この最大値と最小値は、自己組織化マップの提案者の文献を参考として決定した[1)]。以上の 3×3 の条件組合せについて、提案法で用いる乱数の影響を評価するため 10 回の実験を行った。なおその他のパラメータは、$K = 25, T = 100000, S = 100$ とした。

10 回行った実験で、得られた不適合度の値を図 9 に示す。この数値例の場合設定条件の範囲では、初期近傍半径 $\sigma_0^{(0)}$ は小さいほど、初期学習率係数は大きいほど不適合度は改善することが分かる。また、収束過程で改善に失敗し、異常に悪い結果を与える場合が散見されていることが分かる。適用に当たっては、複数回、乱数初期値を変更し、不適合度での最良結果を選ぶ必要がある。

得られたマップのうち、9 通りのそれぞれのパラメータ設定で、乱数の種が 1 の実験より得られたマップ図?? に示す。近傍半径が大きくなると、ノード間の距離が必要以上に離れてしまう傾向がある。

6. カード利用履歴への適用

本節では、あるクレジットカードの利用履歴を対象とした適用事例を与える。サンプルセットの変量を表 1 に示す。各変量は金額であるので、対数変換して用いる。各変量の過去 6ヶ月の月々の金額を並べて当月のその利用者の特徴量 (11×6 次元) としている。なお、月を並べる順番を利用と返済のバランスを示すキャッシュフローで整序したものとしてベクトル化している。個人のサンプリングに当たっては、破産に至った利用者の比率が高くなるように抽出を行った。参照ベクトルの表現法を表 11 に示す。 $\alpha = 0.5, \sigma = 3.0, K = 100, T = 100000, S = 50$

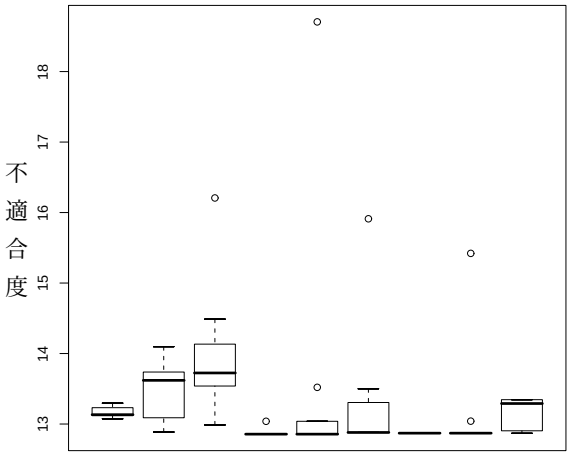


図 9 パラメータを変えて得られた不適合度

で実験を行う。標準の SOM により得られたマップを図 12、可変自己組織化マップにより得られたマップを図 13 に示す。また、その不適合度を比較したグラフを図 14 に示す。標準の SOM の結果では、上部にショッピング利用の顧客、下部にキャッシング利用の顧客の領域があり、右端に休眠顧客のノードがある。提案法では、上記それぞれは、左のノードクラスター、右のノードクラスター、中央上方のノードに対応し、それぞれ独立したノードクラスターとして、分離して表現される結果が得られた。

7. おわりに

標準の自己組織化マップを、任意形状の格子点集合の利用を許容するように拡張することで、高次元特徴量データのもつ自然な位相構造を低次元マップとして表現する可変自己組織化マップを提案した。この方法により、標準の自己組織化マップと比べ、等しいノード数で不適合度を改善することができるとともに、異なった特徴量類型が非連結なノードクラスターと表現されるなど、位相構造に合ったマップが生成できることを数値例や適用事例から示した。

安定して収束が得られる収束パラメータの設定条件の指針の検討や、今回は検討しなかっ

表 1 クレジットカード利用履歴の変量

変量名	説明
SP1 回払い	返却回数が 1 回の月払いの SP 利用金額
SP リボ払い	リボルビング払いの SP 利用金額
SP ボーナス払い	ボーナス一括払い、年 N 回払いを まとめた SP 利用金額
SP 分割払い	上記の SP 利用以外の支払い形態を まとめた SP 利用金額
SP 利用残高	当月における SP 利用残高金額
CS1 回払い	返却回数が 1 回の月払いの CS 利用金額
CS リボ払い	リボルビング払いの CS 利用金額
CS 分割払い	上記の CS 利用以外の支払い形態を まとめた CS 利用金額
CS 利用残高	当月における CS 利用残高金額
入金元金	当月における入金元金
入金手数料	当月における入金手数料

注, SP:ショッピング, CS:キャッシング

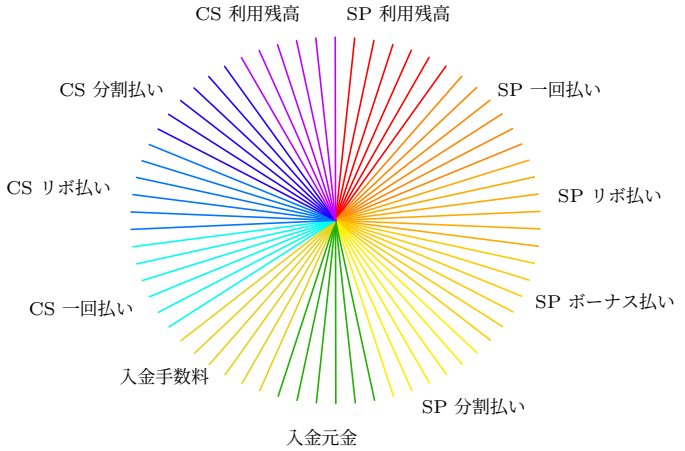


図 11 参照ベクトル

た計算量の低減などが今後の課題となる。

参 考 文 献

1) T. コホネン: 自己組織化マップ, シュプリンガー・フェアラーク東京 (1996)
2) T.W.Anderson: An introduction to Multivariate Statistical Analysis, Wiley (1984)
3) 高根芳雄: 多次元尺度法, 東京大学出版会 (1980)
4) Sam T. Roweis and Lawrence K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science , Vol.290 , pp.2323-2326 (2000)
5) Joshua B. Tenenbaum, Vin de Silva, and John C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science , Vol.290 , pp.2319-2323 (2000)
6) 福井健一, 斎藤和巳, 木村昌弘, 沼尾正行: 自己組織化ネットワークによる動的クラスタの可視化編纂, 人工知能学会論文誌 , 23 巻 5 号 , SP-E , (2008)
7) 関庸一, 長井歩, 石原純一郎, 渡辺亮: 自己組織化マップによる行動履歴の類型化 –クレジットカード利用履歴を利用したキャッシング移行予測–, 日本経営工学会誌 , 57, 5, 404-412 , (2006)
8) 五反田剛, 石井良和, 原健一郎, 関庸一: SOM によるファン層の解析に基づく CD 購買予測モデルの作成, オペレーションズ・リサーチ ,52, 2 , 87-93 , (2007)
9) Yoichi Seki, Eiji Okawara, State Diffusion Model on SOM map based on Multinomial Logistic Model, 4th World Conference of the International Association for Statistical Computing , Dec , 7(5-8) , Yokohama , japan , pp.1391-1396 (2008)
10) 今村弘樹, 藤村誠, 黒田英夫: k 近傍の最大距離に基づくノイズにロバストな自己組織化マップに基づくクラスタリング手法, 情報メディア学会誌 ,62(10) , pp.1618-1623(2008)
11) 沈侃, テキヒ, 北英輔: 自己組織化マップを用いた進化的アルゴリズムについて, 情報処理学会 , 研究報告 , MPS-62 , (2006)

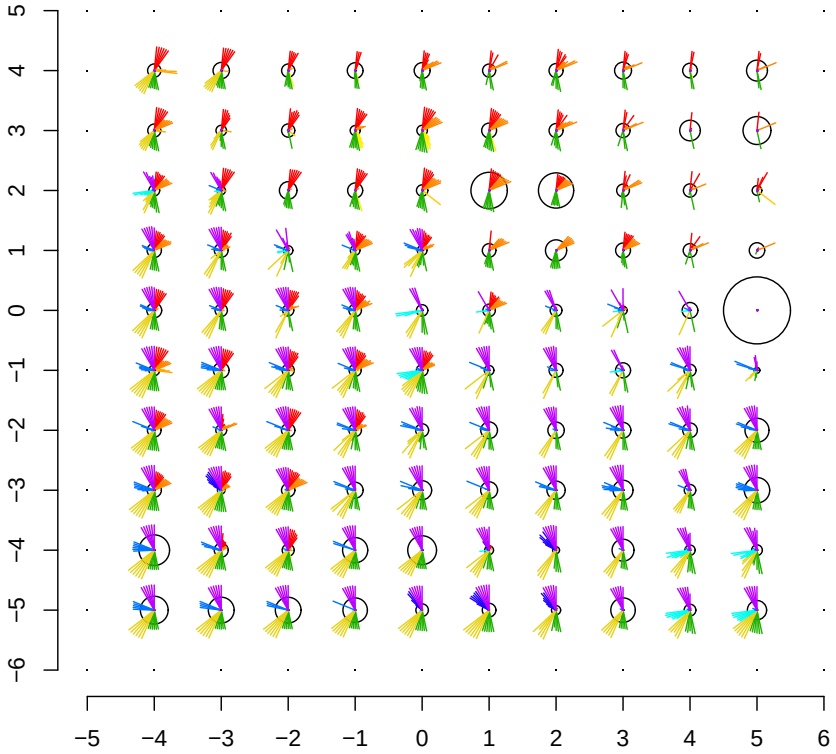
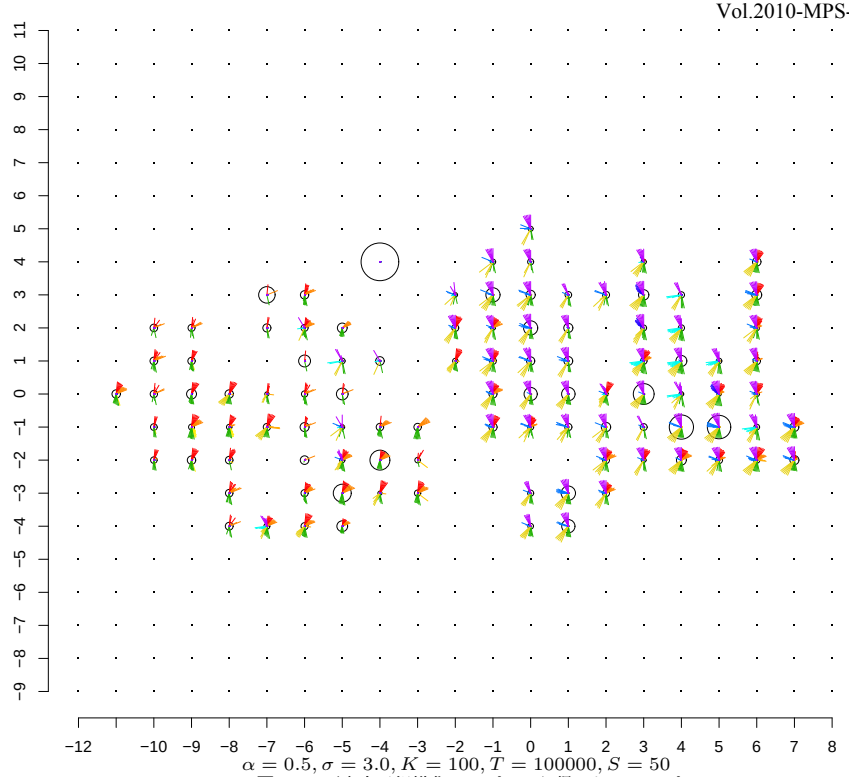


図 12 通常 SOM により得られたマップ
$\alpha = 0.5, \sigma = 3.0, K = 100, T = 100000, S = 50$



$\alpha = 0.5, \sigma = 3.0, K = 100, T = 100000, S = 50$

図 13 可変自己組織化マップにより得られたマップ

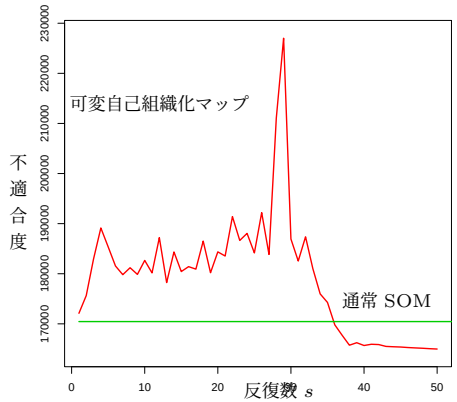


図 14 クレジットカードの利用履歴をサンプルとした場合の不適合度比較

正　誤　表

- 1 ページ左　11 行目　(誤)adapted　to　→ (正)using
- 1 ページ左　16 行目　(誤)adapted　to　→ (正)using
- 5 ページ左　9 行目　(誤) 近傍半径 $\sigma_0^{(0)}$　→ (正) 初期近傍半径 $\sigma^{(0)}$
- 5 ページ左　11 行目　(誤) 学習率係数 $\alpha_0^{(0)}$　→ (正) 初期学習率係数 $\alpha^{(0)}$
- 5 ページ左　16 行目　(誤)$\sigma_0^{(0)}$　→ (正)$\sigma^{(0)}$
- 5 ページ左　5 節最終段落　削除