社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＩＣＳ－140（4）
2005／8／1

自己組織化マップによる教師情報を用いた可視化アーキテクチャの提案

時系列医療データの可視化を例に

福井　健一[†]　斉藤　和巳[††]　木村　昌弘[†††]　沼尾　正行[†]

† 大阪大学産業科学研究所
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
†† NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
††† 龍谷大学理工学部電子情報学科
〒 520-2194 大津市瀬田大江町横谷 1-5
E-mail: †{k-fukui,numao}@sanken.osaka-u.ac.jp, ††saito@cslab.kecl.ntt.co.jp, †††kimura@rins.ryukoku.ac.jp

あらまし　本稿では，自己組織化マップ (SOM) によるクラスタリングおよび類似クラスタが低次元マップ上の近隣に配置される特性を十分活かしつつ，マップの表示法に教師情報を利用することで，クラス分類に関与する系列内クラスタを示唆するマップを構築する可視化アーキテクチャを提案する．本提案法は，SOM による勝者ニューロン系列の学習と，それらを入力としクラスラベルを用いてパーセプトロンによる識別関数の学習の 2 段階から成り，得られた結合荷重をマップにリバースすることで可視化される．実際の時系列医療データを用いた評価実験では，単純なクラス要素数比による表示法と比較し，可視化による定性的評価および分類性能による定量的評価を行った．

キーワード　自己組織化マップ，教師情報，パーセプトロン，視覚的データマイニング，時系列，医療データ

Architecture for Visualization Using Teacher Information based on SOM

Empirical Studies on Visualizing Time Series of Medical Data

Ken-ichi FUKUI[†], Kazumi SAITO[††], Masahiro KIMURA[†††], and Masayuki NUMAO[†]

† The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567-0047
†† NTT Communication Science Laboratories
2-4 Hikaridai, Seika, Kyoto 619-0237
††† Department of Electronics and Informatics, Ryukoku University
1-5 Yokotani,Seta Oe-cho, Otsu, Shiga, 520-2194
E-mail: †{k-fukui,numao}@sanken.osaka-u.ac.jp, ††saito@cslab.kecl.ntt.co.jp, †††kimura@rins.ryukoku.ac.jp

Abstract　In this paper, we propose an architecture for visualization that constructs the map suggesting the cluster in a sequence that involves in classification by utilizing the teacher information for the display method of the map, while making the best use of the characteristics of Self-Organizing Maps (SOM) that are to create clusters and to arrange the similar clusters near within the low dimensional map. This proposal method consists of 2 learning phases, firstly the sequence of the winner neurons are obtained by SOM, secondly connectivity weights are obtained by perceptron, finally the map is visualized by reversing the obtained weights into the map. In the experiments using time series of real-world medial data, we evaluate the visualization and classification performance by comparing the display method by the ratio of classes.

Key words　Self-Organizing Maps, Teacher Information, Perceptron, Visual Data Mining, Time Series, Medical Data

1. はじめに

Kohonenの自己組織化マップSOM(Self-Organizing Maps) [1]は，入力データの類似度を反映したマップを生成するニューラルネットの教師なし学習法である．SOMは，類似したサンプル群をニューロンに対応する参照ベクトルとして汎化する．ここで，あるサンプルに対する最近参照ベクトルは勝者ニューロンと呼ばれ，サンプル群は勝者ニューロンによってクラスタリングされる．そして，SOMではニューロン間に予めトポロジーが定められているため(可視化されるように通常は2次元か3次元の低次元に設定する)，類似した参照ベクトル群が低次元のトポロジー空間内で互いに近くに配置されるように参照ベクトルを更新する．この特徴は，高次元データを低次元へ埋め込む古典的な方法である多次元尺度法[2]や，Sammon's Mapping [3]のように，全サンプルを互いの類似度に基づいて低次元に埋め込むのとは異なり，参照ベクトルにより汎化して埋め込むため，大規模なデータに対する視覚的なデータマイニングに適した手法と考えられる．

通常のSOMは教師なし学習であるため，各サンプルの属するクラスのラベルが教師情報として与えられていたとしても，この情報を有効に利用することができない．そこで，教師ありSOMとして，KohonenのLVQ-SOM [1]や，他にも[4]〜[6]などが提案されているが，いずれも参照ベクトルの学習に教師情報を考慮することで，クラス分類性能の向上を試みている．教師あり学習にすることで分類性能の向上は図れるものの，教師なし学習の本質である潜在クラスタを発見するという意味合いは薄れる．

本研究では，クラスタリングによりデータの素性を知るという意味で教師なしの通常SOMを基盤として，そして教師情報を用いて各クラスタの意味づけを行うことを考える．ところで，SOMの表示方法としては，クラスラベルが与えられていない場合，クラスタに属するサンプル数による濃淡で表す場合が多い[7]．クラスラベルが与えられている場合であっても，各クラスタ内で多数決によりクラス代表ラベルを決定しそれらを表示する[6], [8]に留まっており，教師情報を充分に活かしているとは言えない．

そこで本稿では，SOMによる勝者ニューロンの学習と，パーセプトロンによるニューロン結合荷重の学習の2段階の学習から成る可視化アーキテクチャを提案する．時系列データなどデータが系列として与えられる場合，各系列データは勝者ニューロンの発火系列として得られる．この特性と教師情報を用いて，パーセプトロンにより勝者ニューロンの発火系列を各インスタンスとして識別関数を構成する．そこから得られる結合荷重をマップ上のノードにリバースする事で各クラスタはノード値によって意味付けされる．本手法は，系列データ内の特定の期間に潜在するクラスタの内でクラス分類に関係するクラスタを示唆する意味を持つ可視化が成されると期待される．

実際の時系列医療データを用いた評価実験では，単純にクラスの要素数比によって表示する方法と比較し，可視化による定性的評価および分類性能による定量的評価を行った．

2. 可視化アーキテクチャ

2.1 SOM学習モデル

まず，Kohonenの自己組織化マップSOM(Self-Organizing Maps)について概説する．SOMは，入力データの類似度を反映したマップを生成するニューラルネットの教師なし(unsupervised)学習法である．入力データをN個のV次元ベクトル$\mathbf{x}_n = (x_{n,1}, \cdots, x_{n,V}), (n = 1, \cdots, N)$とする．このマップは，通常2次元格子等間隔に配置されたニューロン(ノード)群で構成され，各ニューロンには入力データと同じ次元の参照ベクトル$\mathbf{m}_j$が割り当てられる．SOMは，次式に基づきニューロンの参照ベクトルを再帰的に更新する．

$$\mathbf{m}_j^{new} = \mathbf{m}_j + h_{c(\mathbf{x}),j}[\mathbf{x} - \mathbf{m}_j]. \quad (1)$$

$$c(\mathbf{x}) = \arg\min_i \|\mathbf{x} - \mathbf{m}_i\|. \quad (2)$$

ここで，係数$h_{c(\mathbf{x}),j}$は近傍関数を表し，その最初の添え字$c(\mathbf{x})$は式(2)で求める勝者ニューロンのインデックスを表す．

近傍関数$h_{c(\mathbf{x}),j}$には，多くの場合，以下のようなガウス分布に基づく定義が用いられる．

$$h_{c(\mathbf{x}),j} = \alpha \exp\left(-\frac{\|\mathbf{r}_j - \mathbf{r}_{c(\mathbf{x})}\|^2}{2\sigma^2}\right). \quad (3)$$

ここで，$\mathbf{r}_j$は2次元格子上に配置された第jニューロンの座標を表し，αとσは学習を制御するパラメータで，ある値から徐々に単調に減少させる戦略が通常よく用いられる．

一方，SOM学習モデルの目的関数は次式で表される[9]．

$$E_{SOM} = \sum_{i=1}^{M} \sum_{\mathbf{x}_n \in C_i} \sum_{j=1}^{M} h_{i,j} \|\mathbf{x}_n - \mathbf{m}_j\|^2. \quad (4)$$

ただし，MはSOMの総ニューロン(ノード)数を表し，C_iは$c(\mathbf{x}) = i$となるサンプル集合を表す．

2.2 SBSOM

通常のSOMのマップには軸に絶対的な意味を持たない．我々は以前，マップの直感的解釈性向上のため系列データに対してSOMの特性を保ちながらある軸方向に系列として配置されるようにSOM学習を修正したSeqence-Based SOM(SBSOM)を提案した[8]．本稿での提案法は，系列データに対する可視化アーキテクチャであるため，SBSOMが適用可能である．ただし，本稿で提案する可視化層の学習にとって，SBSOMを用いる事とは依存しないことに注意しておく．

SBSOMでは，入力ベクトルを$(\mathbf{x}_n, t_n)$，参照ベクトルを$(\mathbf{m}_j, s_j)$と拡張する．ここで，t_n，s_jは系列インデックス(タイムタグなど)であり，両者とも同様に離散化されているものとする．そして，勝者ニューロン選択における距離定義を以下のように修正する．

$$c(\mathbf{x}) = \arg\min_j \delta(t_n, s_j) ||\mathbf{x}_n - \mathbf{m}_j||. \quad (5)$$

$$\delta(t_n, s_j) = \begin{cases} 1 & if \quad t_n = s_j, \\ \infty & otherise. \end{cases} \quad (6)$$

例えば，ニューロンの系列インデックスを同列ニューロンは等

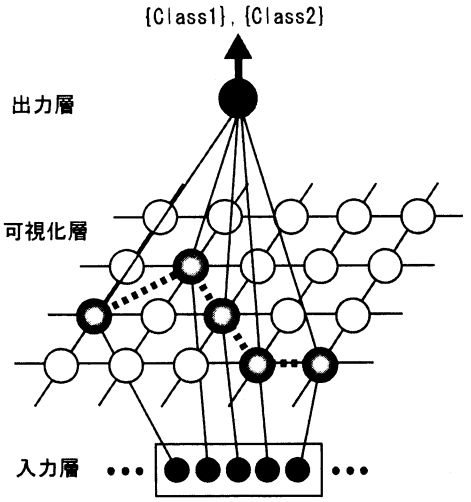

図 1　可視化アーキテクチャ

しくし，行方向に昇順 (または降順) に設定すれば，出力されるマップは横軸がその系列順の意味を持つことになる．ニューロン間の近傍関数については修正していないので，系列の前後のデータの影響受けて，行方向の近傍にも類似したサンプル群が集まる．よって，SBSOM は SOM の特性を活かしながら，マップの軸に系列順の意味を付加している．

2.3　可視化層の学習

本稿で提案する教師情報を利用した可視化アーキテクチャを図 1 に示す．本稿では前提条件として，2 クラス問題のみを取り扱い，また各データは系列として与えられるものとする．まず，K 個の系列データのインデックスを k とし，第 k データの系列数を $n(k)$ とすると，第 k 系列データは $\{(\mathbf{x}_{n(r)}^{(k)}, t_{n(r)}^{(k)}) : r = 1, \cdots, k\}$ と表される．ここで，$\sum_{k=1}^{K} n(k) = N$ となる．可視化アーキテクチャは以下の 3 つのステップから成る．

S1.　$\forall r, k$ $(\mathbf{x}_{n(r)}^{(k)}, t_{n(r)}^{(k)})$ をそれぞれ (SB)SOM の入力データとして学習を行う．その結果，各系列データは可視化層の勝者ニューロンの系列として得られる (図 1 の可視化層は，ひとつの系列データのみ表している)．

S2.　次に，ラベル情報を用いて勝者ニューロンの系列を入力として単純パーセプトロンによりクラス分類をする識別関数を構成する．具体的には，パーセプトロンへの入力データの第 k 系列データの第 i(ノード) 成分 $(i = 1, \cdots, M)$ を次のようにする．

$$y_{k,j} = \begin{cases} 1/n(k) & if \quad \exists r\ c(\mathbf{x}_{n(r)}^{(k)}) \in C_j, \\ 0 & otherwise. \end{cases} \tag{7}$$

そして，ニューロン結合荷重を $\mathbf{w} = (w_1, \cdots, w_M)$ とすると，識別関数は $\hat{z}_k = \mathbf{w} \cdot \mathbf{y}_k$ と表される．$z_k = \{-1.1\}, (k = 1, \cdots, K)$ をクラスラベルとし，次の目的関数を最小化する．

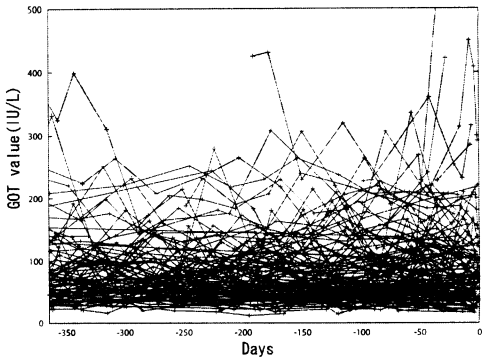

図 2　実際の GOT 値
横軸は IFN 投与日 (右端が投与日) からの相対的な日数を表している．

$$E_{p,t} = \sum_{k=1}^{K} (z_k - \hat{z}_k)^2 + \beta \sum_{j=1}^{M} w_j^2. \tag{8}$$

ここで，第 2 項は Weight Decay(注1) と呼ばれ過学習を抑制する効果がある [10]．係数 β は，Weight Decay の効果を調節するパラメータである．

S3.　最後に，得られた結合荷重 w_j を可視化層の各ノードの濃度なり色相なりに設定することでマップは可視化される．

この可視化アーキテクチャの利点としては，パーセプトロンの観点からは，ブラックボックスになりがちな識別関数の解釈を可視化層に引き出す事になり，SOM の観点からは潜在クラスタの中でクラス分類に関与する部分に焦点を当てた可視化がなされる，と期待される．

3.　実　　験

提案法の評価に，肝炎患者の医療データを用いて，可視化による定性的評価および分類性能のよる定量的評価実験を行った．

3.1　肝炎患者データについて

まず，本実験で使用したデータについて説明する．実際の病院のカルテから収集された C 型肝炎患者 137 人 1 年間分の血液検査の時系列データを用いた．今回用いた検査項目は肝炎と何らかの関係があると考えられている GOT,GPT,TP,ALB,T-BIL,D-BIL,I-BIL,TTT,ZTT,CHE,T-CHO の 11 項目である．患者毎に検査間隔に差はあるものの，およそ月に 1 回程度の頻度であった．一例として，図 2 には全患者の GOT 値をプロットしたグラフを示すが，このように医療データは変化が大きくかつ多次元 (他項目) のデータである．横軸は IFN 投与日から何日前か相対日数を表している．

3.2　問題設定および目標

近年，肝炎の治療薬としてインターフェロン (IFN) が用いられている．しかしながら，IFN は高価であり，また副作用が強い

(注1)：結合荷重の 2 乗和のようなペナルティ項を目的関数に加えることにより，不要な結合荷重が大きくなり過ぎないように抑制する方法である．

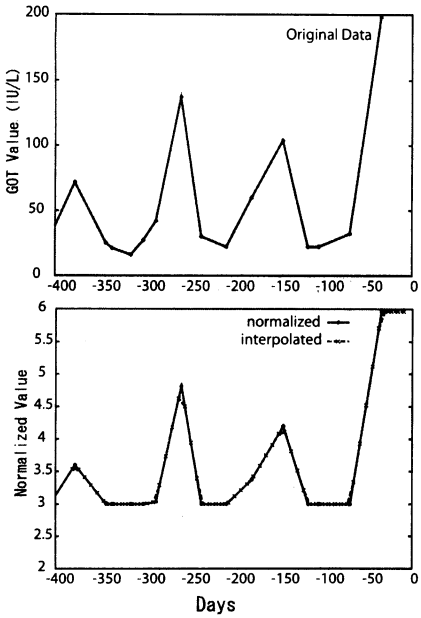

図 3　前処理:規格化・線形補間の例

反面，治癒率は平均するとおよそ 30%とそれ程高くはない(注2)．ゆえに，IFN 投与以前の検査結果から患者への IFN の有効性を予測することは，患者に対する肉体的，精神的，コスト負担という点において重要な課題である．本データは，HCV-RNA 検査(注3)により IFN 投与前後での肝炎ウィルスの存在有無によって，予め正例 (55 例)・負例 (82 例) の 2 クラスに分類されている．

本実験の目標は，IFN が有効な (または有効でない) 潜在的かつ特定期間における患者グループを示唆するマップを構築することである．パーセプトロンより学習した結合荷重による表示法と単純に各クラスタに含まれる正例・負例の割合による表示法とを比較して，提案手法の有効性を検証する．

3.3 前　処　理

データの前処理としては，SOM において類似度を求める際に属性 (検査項目) 間の意図しないバイアスを避けるために，各属性は医師により提示された指標 [11] を元に規格化を施した．平常値範囲内の値を 3 とし，異常に低い値・高い値を足切り頭切りし，1 から 6 までの連続値内に納めた．また，患者毎に検査間隔は不定であるため，1 週間単位で線形内挿補間を行ってデータの離散点を一致させた．(図 3)

3.4 可視化結果

次に，SBSOM(15 × 52 ノード) による同じ学習結果を用いて，マップ上に表すノード値をそのノード (クラスタ) に属する (正例数)/(総数) による比とした結果 (図 4) と，パーセプトロンにより学習した結合荷重により表示した結果 (図 5) をそれぞれ示す．ただし，比による方法は，クラスタの要素数に応じた信頼度を重みとして乗じてある．横軸は IFN 投与日からの相対的な週数を表している．縦軸は通常の SOM 同様絶対的な意味は持たず，クラスタ間の類似関係を保存するような相対的な意味合いしかない．各マップには，例として，ある正例患者と負例患者の勝者ニューロン系列 (患者のパスと呼ぶことにする) を共に図示した．患者のパスが値の高いノードまたは低いノードを通過している場合，そのクラスタに属する他の患者の症例と見比べるなど視覚的データマイニングに利用するものと考える．

検査データの少ない 30 週以前については，要素数が 1 または 2 のクラスタが多いため，比による表示法では，それらのノード値は 0 または 1 になりがちになり，前半期間の少数派のクラスタにバイアスがかかってしまっている．それに対して，結合荷重による表示では，そのような偏りは見られず，全期間に渡ってにバランス良く表れている．

3.5 分類性能評価

正しく正例・負例を分別するようなノード値を表しているかを検証するために，各患者のパスを元に正・負例の正答率を算出し評価を行った．ここで正答率とは，正例を正例と判別または負例を負例と判別した割合を表す．比による方法の場合，患者のパスのノード値平均値が 0.5 以上ならば正例，未満ならば負例と判別し，また提案法の場合，パーセプトロンの識別関数の出力の正負を判別基準とした．

表 1 に 10-fold cross validation で評価した結果を記載してある．比による方法では，試験例の正答率の最大値と最小値の平均は 47.53%であった．これはランダムとほぼ同等の結果であるため，未知の患者に対応するための汎化された分別能力は全くないと言える．提案手法では，訓練例において正答率はほぼ 100%，試験例においては 64.29%であった．しかし，試験例に対する正答率の分散が大きいため，良い結果とは良い難い．原因として過学習している事が考えらるが，式 (8) における β を変えて実験を試みたが，その効果は得られなかった．この原因については，さらなる調査が必要である．表 1 は $\beta = 0$ の時の結果である．

手法	比による方法	提案手法
判別基準	平均値	識別関数
訓練例	79.78 ± 4.78	99.60 ± 0.40
試験例	47.53 ± 16.76	64.29 ± 21.43

表 1　10-fold cross validation の結果

参考までに，本データと同じ期間・属性を用いた他の研究ついては，属性値の勾配を捉えるような述語を設定し，帰納論理プログラミング (ILP) を用いて分類規則を抽出した研究 [11] では，試験例における平均正答率は 67.2% であった．また，離散化して補間した検査値から患者間のオーバーラップ期間を最小単位とする前処理をし，それらの組み合わせで分類規則を最小記述長原理 (MDL) を利用して構成した研究 [12] では，平均正答率は 73.03% であった．これらからも，予測の難しいデータである事が伺われる．

(注2)：http://www.e-chiken.com/shikkan/c-kanen.htm
(注3)：C 型肝炎ウィルスの遺伝子の一部を，PCR 法という遺伝子増幅技術によって増やし，ウィルスが存在するかを直接確認する検査法で，信頼性の高い検査法として知られている．

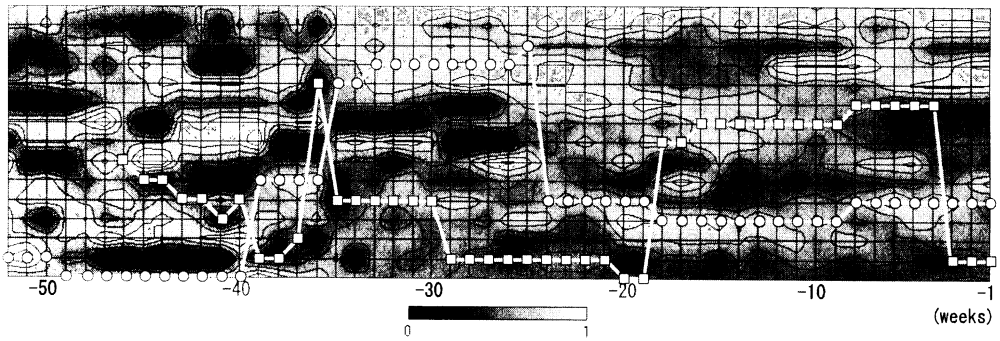

図 4　比による表示
ある二人の患者 (○が正例患者，□が負例患者) のパスを共に図示してある

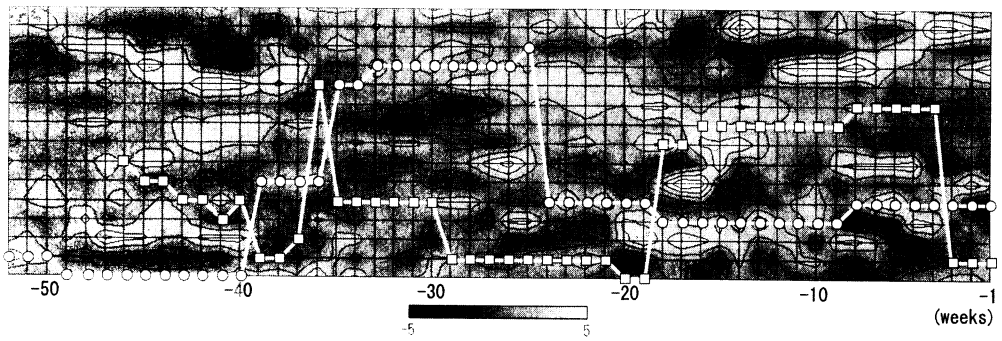

図 5　結合荷重による表示

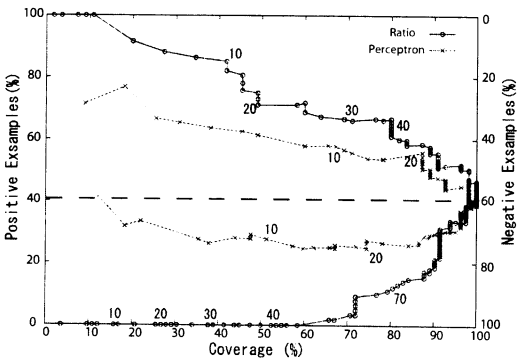

図 6　クラスタセット
上半分はノード値の降順，下半分は昇順のクラスタセットの値をそれぞれプロットしている．グラフ中の数字は順位を表している．

3.6　クラスタセット

次に，どれだけ多くの正例 (または負例) を含むクラスタをマップで示唆できているかを検証するために，クラスタセット (クラスタの集合) を考える．図 6 には，各ノード値の降順および昇順にクラスタをピックアップしていき要素数を増やしていった時の正例の被覆率 (全正例中クラスタセットに含まれる正例の割合) を横軸に，クラスタセット中の正例の割合を縦軸に表している．当然，降順の場合は少ないクラスタ数で多くの正例を含み，かつ負例はなるべく含まない方が良いと考えられる．昇順の場合はその逆となる．

まず，図 6 の上半分に着目すると，パーセプトロンによる方法では，最上位付近に含まれる正例の割合は比による方法に比べて低くなるものの，被覆率は高く上位 20 クラスタで，比による方法のおよそ 45 クラスタ分とほぼ同等のクラスタセットが得られている．昇順についても同様の傾向が伺われ，提案法はノード値の上位 (下位) 少数のクラスタでより多くの正例 (負例) を含んでおり，マップが示唆するクラスタとしてより有益なものとなっていると言える．

4.　おわりに

本稿では，教師情報を持つ系列データに対して，パーセプトロンを利用して SOM の結果を表示する可視化アーキテクチャを提案した．本提案法の利点は，SOM によるクラスタリングおよび類似クラスタが低次元マップ上の近隣に配置される特性を十分活かしつつ，マップの表示法に教師情報を利用することで，クラス分類に関与する系列内クラスタを示唆する可視化が成される所にある．現実の時系列医療データを用いた初期実験では，単純に正・負例の比により表示する方法と比較して，提

案法は次の 3 点において優れていることを確認した.

- データが少なく少数クラスタが多い期間であっても偏りなく表示する事ができている.
- 充分とは言えないが汎化された分別能力が得られていることを確認した. 前処理の再検討, 属性値の勾配や時間伸縮性や考慮, 過学習の抑制など考える必要はあるが, 分類に関与するクラスタを示唆する可視化が成される可能性を事を示すことができた.
- ノード値の上位 (下位) 少数のクラスタでより多くの正例 (負例) を含んでおり, マップが示唆するクラスタとしてより有益なものとなっていると言える.

今回は, ノード値の学習にパーセプトロンを用いたが, 与えられたデータから, ある情報 (関係式) を元にして, 元々の分布を再構成する方法として知られる最大エントロピー法 (Maximum Entropy Method) が適しているかもしれない. また, 教師情報を用いての可視化の有用性を確かめるための別の角度からの実験も必要である.

文　献

[1] T. Kohonen: "Self-Organizing Maps", Springer-Verlag, Heidelberg (1995).

[2] J. B. Kruskal and M. Wish: "Multidimensional scaling", Number 07-011 in Paper Series on Quantitative Applications in the Social Sciences (1978).

[3] J. Sammon: "A nonlinear mapping for data structure analysis", IEEE Transactions on Computers, **c-18**, pp. 401–09 (1969).

[4] R. Hecht-Nielsen: "Counterpropagaton networks", IEEE Frist international Conference on Neural Networks, pp. 19–32 (1987).

[5] B. Fritzke: "Growing cell structures: A selforganizing networks for unsupervised and supervised learning", Neural Networks, **7**, pp. 1441–1460 (1994).

[6] 福田, 斉藤, 松尾, 石川: "教師情報を導入した som 学習モデル", Technical Report NC2003 - 142, 信学技報 (2004).

[7] T. Kohonen, S. Kaski, K. Lagus, J. Salojrvi, J. Honkela, V. Paatero and A. Saarela: "Self organization of a massive document collection", IEEE Transaction on Neural Networks, **11(3)**, pp. 574–585 (2000).

[8] K. Fukui, K. Saito, M. Kimura and M. Numao: "Sbsom: Self-organizing map for visualizing structure in the time series of hot topics", Proc. Joint Workshop of Vietnamese Society of AI, SIGKBS-JSAI, ICS-IPSJ, and IEICE-SIGAI on Active Mining, pp. 19–24 (2004).

[9] T. Kohonen: "Comparison of som point densities based on different criteria", Neural Computation, **11**, pp. 2081–2095 (1999).

[10] S. Hanson and L. Pratt: "Comparing biases for minimal network construction with back-propagation", Advances in Neural Information Processing Systems 1, pp. 177–185 (1989).

[11] 佐藤: "Ilp を用いた医療データからの知識発見", 東京工業大学大学院情報理工学研究科 計算工学専攻 修士学位論文 (2005).

[12] 本山, 市瀬, 沼尾: "間隔不定な時系列データからの知識発見", 人工知能学会研究会資料 SIG-KBS-A405-05, pp. 27–32 (2005).