# k次伝播SOMによるデータ分類

MIURA, Takao / SHIOYA, Isamu / 塩谷, 勇 / YANAGIDA, Taqlow / 柳田, 卓郎 / 三浦, 孝夫

(出版者 / Publisher)
一般社団法人情報処理学会
(雑誌名 / Journal or Publication Title)
情報処理学会研究報告.データベースシステム (DBS)
(号 / Number)
67
(開始ページ / Start Page)
457
(終了ページ / End Page)
464
(発行年 / Year)
2002-07-18

データベースシステム 128－61
（２００２．７．１９）

# $k$次伝播SOMによるデータ分類

柳田 卓郎[1]　三浦 孝夫[1]　塩谷 勇[2]
[1]法政大学 工学研究科 電気工学専攻　[2]産能大学　経営情報学部

本稿では，クラスタ化の観点から自己組織化マップ (以下 SOM) の拡張を論じる．説明機能を強化するために近傍学習機構を導入して k 次伝播 SOM (SOM($k$)) を定義し，実験によりその有効性を検証する。またクラスタリング機能の評価を数学的に行うための検定機能を論じ,SOM ($k$) の検定について述べる。

# $k$-propagated Self-Organizing Maps

Taqlow YANAGIDA[1]　Takao MIURA[1]　Isamu SHIOYA[2]
[1]Dept.of Elect. and Elect. Engr.,HOSEI University
[2]Dept.of Management and Informatics, SANNO University

In this investigation, we discuss Self Organizing Maps (SOM) based classifiers with propagation on SOM map. To enhance *explanation* capability, we introduce *neighbor learning* and define $k$-propagated SOM, called SOM($k$). We show experimental results to see how SOM($k$) is useful. The main idea is that we extend SOM by propagating classification as well as parameter learning of SOM.

*Keywords:* Machine Learning, Artificial Neural Network (ANN), SOM($k$), SOM, classifiers

## 1　目的

教師なし学習は広範な分野のクラスタ分析や分類に使用されている [4, 6]。他にも，人工ニューラルネット (ANN) は工業や商業のクラスタリング手法としてうまく使われているが、ANN 学習 [8] に不可欠な機能となる分類能力が非常に望まれている。ANN には見本として作用する際立ったデータ例に依存せず各々のクラスタを表現する傾向があることは周知である。新規データは何がしかの距離尺度が最も類似した見本を元にクラスタ化される。

代表的な方法は、教師無し競合学習と自己組織化マップの２つである。どちらも競合型 ANN に属していて、前者は入力データに対して winner-takes-all 法で競合するニューロン階層構造型である。クラスタ内で勝者ユニットだけが活性化し、活性化したユニットは上位レイヤへ入力パターンを出力する。レイヤ内の関連はひとつのクラスタ内でただ一つのユニットが活性化しているので抑制的に働く。勝者ユニットは特徴あるデータに強く応答し，その結合重みを修正する。最終的に，各々のクラスタがデータ内の規則性を探索するための新たな特徴として見ることができる。その結果は弱い特徴から強い特徴のマップとして表される。

ANN に関する文献における教師無し学習のその他の形式として Kohonen の自己組織化マップ (以下 SOM)[7] がある。SOM は順序格子によるクラスタの平滑効果を使った次元変換と競合学習を合成したものである。SOM では、クラスタリングは数個のユニットがデータに対して競合することによって形成される。データに最も近い重みベクトルを持つユニットが勝者となってインプットデータにより近くなるよう移動し、勝者の重みは最整合ユニットとして修正される。その集合は類似した入力がお互いにより近い位置になろうとすることから自己組織化マップまたは特徴マップと呼ばれる。SOM は脳内の処理に類似していると考えられ,2 あるいは 3 次元空間上に高次元のデータを可視化したりするために有効に使われることがある。

SOM をクラス分類するものと見る時、そのプロセスは学習ベクトル量子化と呼ばれる。WTA アプローチでは近傍はほとんど独立して最新のパターンの距離に更新されるのだが，これで本当に分類されているの

だろうか。本研究では近傍学習の機能をSOMに追加し,分類の結果が良好なクラス分類を形成していることを検証する。本稿において伝播学習をk回に制限したものをSOM($k$)と呼称する。2章で簡単にSOMを概観し,分類するものとしての側面を強調する。その後3章でk次伝播SOM(SOM($k$))によってこの側面を取り戻すことを示し,4章で統計的検定の観点から分類するものの評価法を示す。全ての議論はDNA microarraysデータを使用した実験結果について述べるものである。

## 2 自己組織化マップとクラスタ化

SOMはビジュアル化の手法のひとつであり、その目的は隠れたデータ構造の発見である。そのアイデアは脳内のニューロンは複数のグループにクラスタ化されている傾向があるという事実から成り立っている。グループ内の関連はグループ外のニューロンとの関連より遥かに大きい。SOMはこの部位をシンプルな方法で模倣しようとする。n次元空間はANNの訓練のためのn組の入力ノードとして表される。出力ノードは全ての入力が全ての出力に結合された2次元または3次元空間上の直角格子となる。ANN上の重みは訓練中与えられた入力によって出力ノードの一つだけが興奮する時修正される。入力ベクトルはより近い出力ノードの一つに近づき,そのノードはやがて補強された重みを得ることになる(図1[7])。遠く離れた入力ベクトルは他の出力ノードを興奮させる。

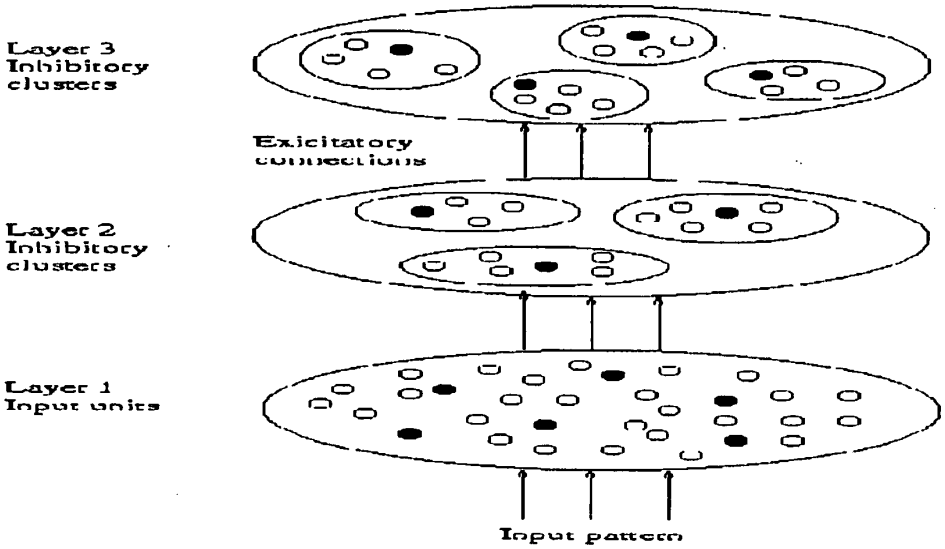


図1: Layers of Self Organizing Map

ANNと時間$t$における入力ベクトルを$p$とすると, 勝者出力ノード$x$を入力と出力をつなぐノードの重みベクトル$w_x$から得たユークリッド距離が最も近いもので表現する。勝者$x$と学習パラメータ$\mu(t,x)$が与えられる事で重みベクトル$w_x(t)$は$w_x(t+1) = w_x(t) + \mu(t,x)(x - w_x(t))$のように更新される。一般的に$\mu(t,x)$は$\alpha \times (1 - t/T)$として与えられる。この式における$\alpha$は学習率係数と呼ばれる数値で、[1]Tは実験間隔である。また近傍y(出力層の勝者に隣接したユニット)が持つ重み$w_y(t)$は $w_y(t+1) = w_y(t) + \mu(t,y)(x - w_y(t))$として少し更新される。本研究では出力層上において,勝者点xを除いた次の8つのノード、すなわち$East(E), West(W), South(S), North(N), NE, NW, SW, SE$を$Nbr(x)$としてあらわす。トポロジー情報は$Nbr(x)$によって与えられる。$\mu$パラメータによって$Nbr(x)$内の近傍ノードは最終的に入力パターンに近い応答をすることから類似した更新を受けることになる。近傍の全てが同じ入力$p$のクラスを持つよう

[1] $0.03 \leq \alpha \leq 0.05$が適当であるとされる

に変えられ, そのプロセスは設定時間経過後か, 出力ノードが安定したときに止まる。2 次元表示が用いられ、興味ある類似したデータ要素はノードの物理的配置を考慮に入れられて近い位置にマップ化される。互いに近いノードは遠く離れているノードよりも強く相互に影響しあう。トポロジマップは単なる近隣の関係を保持したマッピングである。例えば二つの入力ベクトル $p_1, p_2$ に対応する勝者出力ノードが $t_1, t_2$ とする時、$p_1$ と $p_2$ が類似していれば $t_1$ と $t_2$ は近くなるはずである。

SOM による分類は、学習ベクトル量子化 (LVQ) と呼ばれる。LVQ はマップ上の各点にクラスを与えることによって SOM を拡張したもので,SOM との差異はクラス分類のための訓練データの存在である。LVQ マップは SOM と同じく WTA 法の学習によって生成され, 入力ベクトル $p$ はその勝者ノードが割り当てるクラスで分類される。最終的にテストデータを使ってマップの有効性を評価することができる。

LVQ マップが描かれる様子をみてみよう。SOM プロセスにおいてクラス c をもつ入力ベクトル $p$ の勝者点 x を出力層から選択し,$c =< 0, ..., 0, 1, 0, ...0 >$($c$ 番目が 1) を $x$ のクラス分布に加える。訓練データによるプロセスの繰り返しから最終的にクラス分布を得る。LVQ マップの有効性を検証するためにテストデータを使ってクラス分布を集め、訓練データのそれとの比較を行う。勝者が少量でほんの数個のノードとなるならば、これらのノードは検定から除外する。勝者が 5 つ以上現れる時その出力層のノードを $hit-point$ と呼ぶことにする。クラス分布を保持するという点で Kohonen が提案するオリジナルの LVQ[7] とは多少違うものであることに注意してほしい。オリジナルの LVQ では WTA 法によってただひとつのクラスだけが選択される。

**例 1** 流れの説明のために 7 次元のベクトル値で表された 818 件, 18 クラス (0 ～ 17 によって表現する) の DNAmicroarrays information[2][1] に SOM 手法を適用する。本実験においては 818 件のデータの内 300 件を訓練データとして使用し、そこから得られた 7 × 7 の SOM マップを図 2. の左図に示す。このマップ上からは 13 点が hit-point(5 件以上勝者として現れる点) となった。結果を図 2. の右図に示す。□

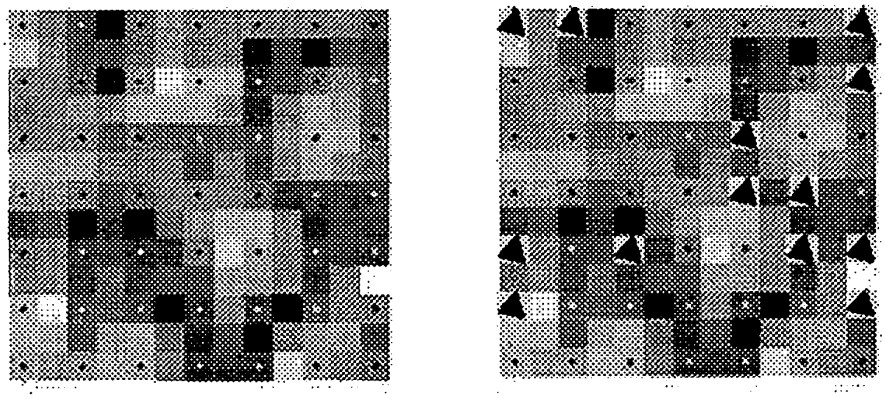

図 2: SOM Map and Hit Points

根本的疑問のひとつは類似した入力は近隣に分類されるのかどうかである。事実、LVQ のトポロジ情報はクラスに対するパラメータ学習能力が LVQ マップ上に存在しないため SOM とは違うもののようだが, 分類能力とトポロジがお互い矛盾するらしいことは周知の事実である。例えば、SOM アルゴリズムの収束は近傍の大きさが 0 になるということであり, マップはトポロジ配列特性でしかなくなってしまう [9]. 他の研究では SOM の拡張はトポロジ情報を保持しないと述べている [3]. また, 得られた LVQ マップは本当に正確であるのかどうかという問題もある。尺度のひとつとして誤分類率を用いる方法がある。しかし数学的に十分とはいえないことから, 本稿は適合度検定の方法をとる。

# 3 $k$ 次伝播SOM

この章では $k$ 回伝播後の近傍のクラス分布を学習するマップ：$k$ 次伝播 SOM(SOM($k$)) 提案することで新しい学習方法を示し，このアプローチの利点と欠点について議論する。今，クラスカテゴリ $\mathcal{C} = \{1, .., n\}$ がありマップ上の全ての点 $x$ がクラス分布ベクトル $x[cnt, c_1, .., c_n]$ と，二つのパラメータ $\mu$ と $\tau$ を持っているとする。$cnt$ パラメータは頻度を，$c_i (i = 1, 2, ...., n)$ はそれぞれ $0.0 \leq c_i \leq 1.0$，$c_1 + ... + c_n = 1.0$ を満たす実数値をとり，クラスメンバシップの相対的分布を示す。例えば $[100, 0.9, 0.1, 0.0, ....., 0.0]$ は 100 個のベクトルのうちクラス 1 を勝者としたクラス分布が 90% であり，クラス 2 を勝者とした分布は 10 % であることを意味する。パラメータ $\mu$ は SOM マップの重みを，$\tau$ は近傍重みを示すものである。

いったん勝者が選ばれると SOM プロセスと同様に近傍のベクトルが更新される。本研究では更新されるクラス情報は伝播方向を付加して保持され、後でパラメータとなってこの変化が次の学習時に勝者の外側に向かって伝播する。この動作を最大 k 回繰り返し行う。ここで注意が必要なのは入力ベクトルは伝播と伝播の間に与えられ，マップ上の全ての動作がこの間に行われる点である。SOM($k$) は以下のような等式であらわすことができる:

$$w_x(t+1) = w_x(t) + \mu(t,x)(x - w_x(t)) + \tau(t,x)\gamma_x(t)$$

$$w_y(t+1) = w_y(t) + \mu(t,y)(x - w_y(t)) + \tau(t,y)\gamma_y(t)$$

$\gamma_x(t)$ は伝播ベクトルであり $x$ は時間 $t$ の関数である。SOM のように勝者 $x$ の $w_x$ ベクトルは $\mu(t,x)(x - w_x(t))$ によって修正後クラス情報と共に更新される。この動作が近傍 y にも適用される。類似した $\tau(t,x), \tau(t,y)$ は出力層上の伝播パラメータである。$\mu, \tau$ はどちらも学習パラメータであることに留意しておく。伝播とは数個のベクトルが出力層内の点の周囲に割り当てられることを指す[2]。考えうる伝播のモデルのうち近傍学習が操作できそうなものとして力学モデルと波動モデルを挙げることができる。本稿では力学的モデルについての議論をするものとして以下に詳細を示す。$\vec{\gamma}$ は $x$ の全ての近傍 $y$ に対する $(w_y - w_x)$ を計算し，$z$ は出力層上の $x$ と隣り合う点の内 $w_x + \gamma$ の勝者となる点をあらわす。ここで伝播方向を $x \to z$，伝播の強さを $|\vec{\gamma}|$ とする。$\vec{\gamma}$ が値なしならば伝播は起こらない。$\vec{\gamma}$ が伝播することによってあたかも別の学習すべき入力が与えられたかのようになり、プロセスはその伝播ベクトルによって動作を続ける。このモデルを力学モデルと呼ぶ理由は多数のベクトルによって表される力の集合のように見えることからくる。LVQ($k$) は SOM($k$) 上に LVQ マップを示す。理論上はマップが安定するまで $k = 0, 1, 2..... \infty$ とプロセスを継続することができる。

**例 2** 実験による $k = 1, 2, 3, 5, 10, 15$ のそれぞれの SOM($k$) 学習の出力と LVQ($k$) マップを重ねたものを図 4 に示す。$hit - point$ は三角印の点か丸印の点で描画されている。図 2 の SOM(0) マップと比較することで伝播がどのように進むのかを見て取ることができる。実際にはこれ以上伝播が進むとさらにトポロジが失われてしまう。図 3 は棒グラフで $k = 1, 2, 3, 5, 7, 10, 15$ の LVQ 情報から得られる SOM($k$) マップ上の $hit - point$ の数を表している。$k = 0$ と $k = 10$ で最も多い 13 個の $hit - point$ をとることがわかる。

# 4 SOM($k$) の評価

SOM($k$) の分類能力の有効性を検定するために統計的な枠組みとクラス分類の結果に対する適合度検定について述べる。テストデータを使って得た出力層各点のクラス分布と訓練データによって得た各点の分布とで検定を行う。統計的検定の見地から、各点のクラス分布が一致すればテストデータは検定に合格することができる。

[2]一方，誤差逆伝播法 (BP) は入力層一個に値を返す

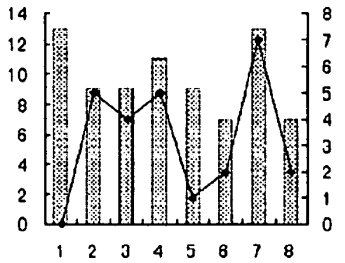


図 3: Number of Hit Points

訓練マップの各点は各々クラス分布を保持していることに注意してもらいたい。一方従来の LVQ で誤分類率によるアプローチが採られているのは、各点がひとつだけクラスを持っているからである。しかし本研究においては訓練データ上に分布を持っているため誤分類率による検定は不適当であると考える。分布とクラスの間のギャップは許容誤差の概念または刈り込みを引き起こす。

ある一点のクラス分布 $< P_1, .., P_n > (P_1 + ... + P_n = 1.0), 1.0 \geq P_i \geq 0.0$ を例にとる。この分布はテストデータに対して利用されるべきでありこのような出力層上の各点確率的なクラス分布を考えているのでこれを新しいクラス分類とみなすことができる。

**例 3** 図 5 に SOM(0),SOM(1),SOM(10),SOM(15) の LVQ 訓練マップの一部を示す。伝播の度合いが高いほどよりマップが拡散しているのが見て取れる。

クラス分布の適合度検定を行う時は各点におけるテストデータのクラスの頻度を数えた結果と前もって訓練データによって定められた分布との間で適合度検定を適用する。$m \times m$ の表で、訓練データの列から分布 $P_i$ を得る。行はテストデータの頻度をあらわす。テストデータを検定するために次の式で表される $X^2$ 値を得る必要がある。

$$X^2 = \sum_{j=1}^{m} \frac{(y_j - Np_j)^2}{Np_j}$$

$N$ がテストデータの各点あたりの件数であることから, 直観的にこの値は分布の期待値と実測値の間の誤差を示す。$\mathcal{X}^2$ 検定を使う事で $\{P_j\}(j = 1, .., m)$ の分布によって表されたテストセットが度の程度有効であるかを評価する。この検定に却下された場合,LVQ マップは分布と一致しないことになる。
本研究で LVQ マップ上の各点で分析が可能であることに注意して違う側面から見ると、この議論は誤分類率からの分析も可能である。しかし, 誤分類率は SOM($k$) 内の数点だけでは信頼性が非常に低く, 不適当である。

**例 4** 実験結果を検定してみよう。$X^2$ 値が $\mathcal{X}^2$ 値よりも小さければその点は信頼できるとして $reliable-point$ とする。図 4 では丸印の点が $reliable-point$ を示す。図 3 で有意水準 0.20 での $reliable-point$ の数を示す。グラフからその点を読みとることができるはずである。

図 6 は LVQ($k$)($k = 1, 2, 3, 5, 10, 15$) 上の各 $hit-point$ における $X^2$ 値と $\mathcal{X}^2$ 値を示す。ここから伝播の度合いが高いとクラス分類の精度が向上することが容易に見て取れる。一方、LVQ(10) は LVQ(15) よりも $hit-point$ が多いことから伝播数が大きすぎると精度を失うと考えられる。

## 5 結び

本研究では $k$ 次伝播 SOM とこの手法によるクラス分類を示した。主たるアイデアは BP-SOM のような従来のアプローチとは違った近傍学習と出力層における学習の伝播である。LVQ に SOM($k$) を適用するた

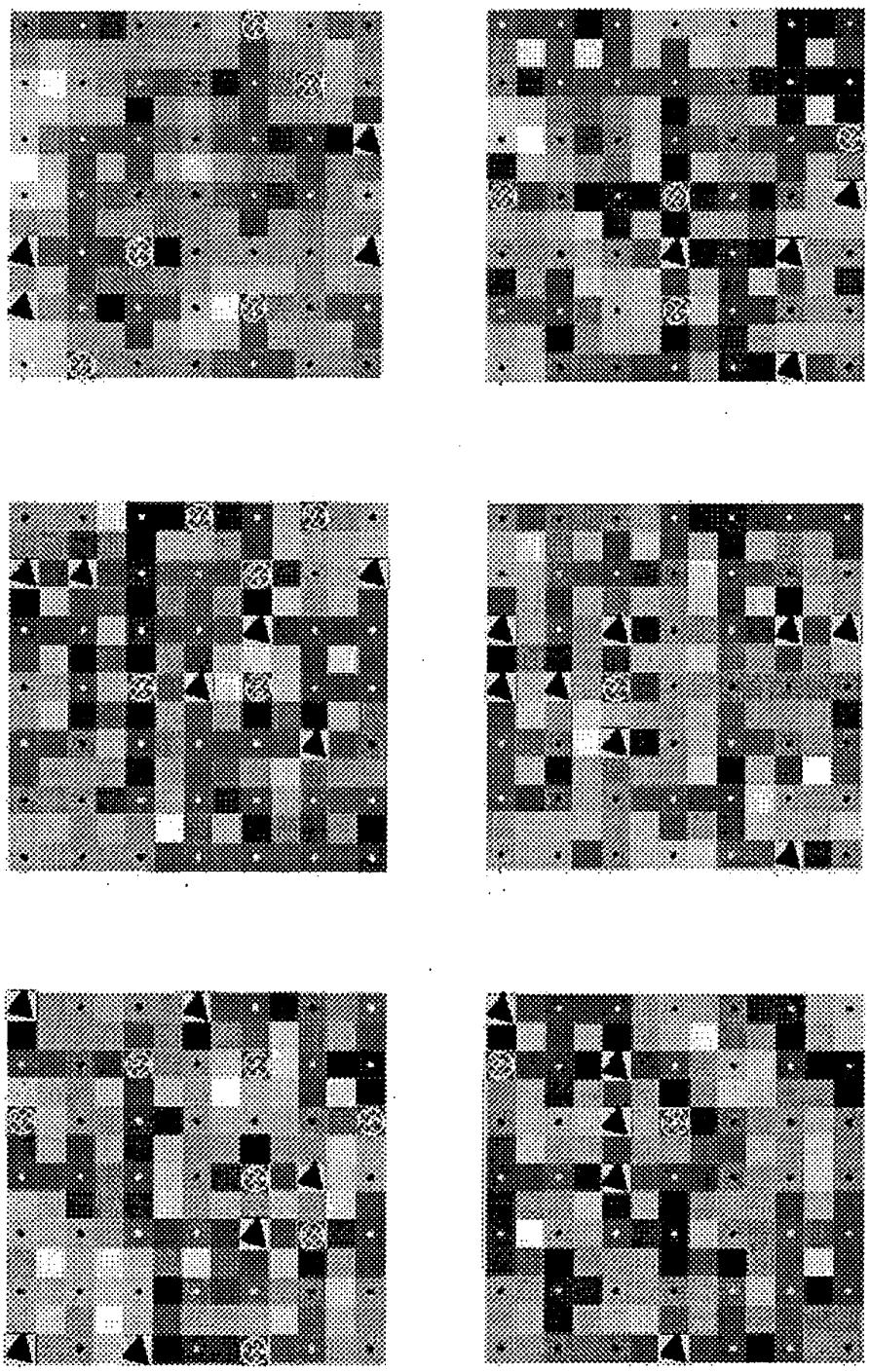

図 4: LVQ(1), LVQ(2); LVQ(3), LVQ(5); LVQ(10),LVQ(15)

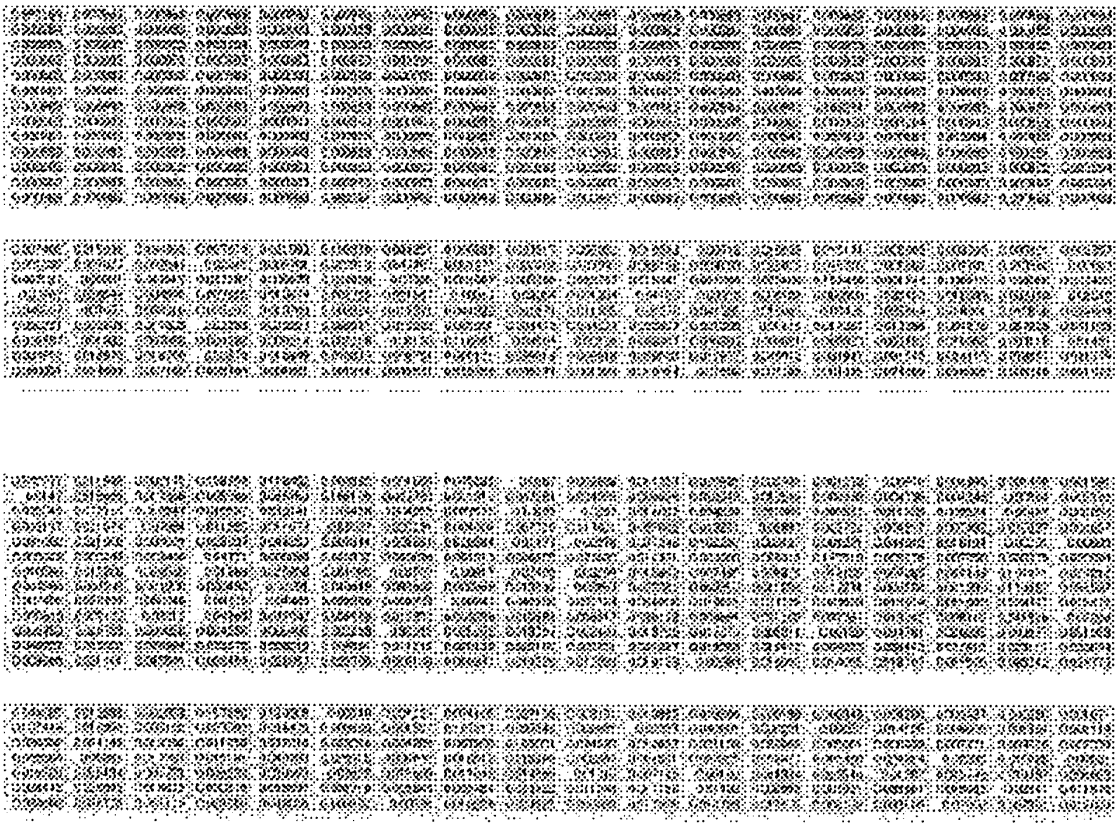

図 5: Training maps of LVQ(0), LVQ(1); LVQ(10), LVQ(15)

めに各点にクラス分布を与え, クラスの伝播を示した。また統計的検定を使う事によってクラス分類の有効性を検定し, この手法によってクラス分類の精度が向上したことを示した。現在新しい応用として波動モデルによる $k$ 次伝播のアプローチを研究中である。

## 参考文献


[1] BioInformatics Units: DNA arrays, Centro National de Investigaciones Oncologicas,http://bioinfo.cnio.es/dnarray/data/

[2] DeRisi et al.: Exploring the Metabolic and Genetic Control of Gene Expression on a genomic Scale, *Science* 278, 1997, pp.680-686.

[3] Desieno, D.: Adding a conscience to Competitive Learning, ICNN, 1989, pp.117-124

[4] Han, J. and Kamber M.: Data Mining - Concepts and Techniques, Morgan Kauffman (2000)

[5] 広津 : 離散データ解析 (1982), Kyoiku Shupan, in Japanese

[6] Kaski, S. and Lagus, K.: Comparing Self Organizing Maps, ICANN, 1996, pp.809-814

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

[8] Weijters, T., Herik, H.J. et al.: Avoiding Overfitting with BP-SOM, *IJCAI*, 1997

[9] Yin, H. and Allinson, N.M.: On the Distribution and Convergence of Feature Space in Self-Organizing Maps, Neural Computation 7, 1995, pp.1178-1187.

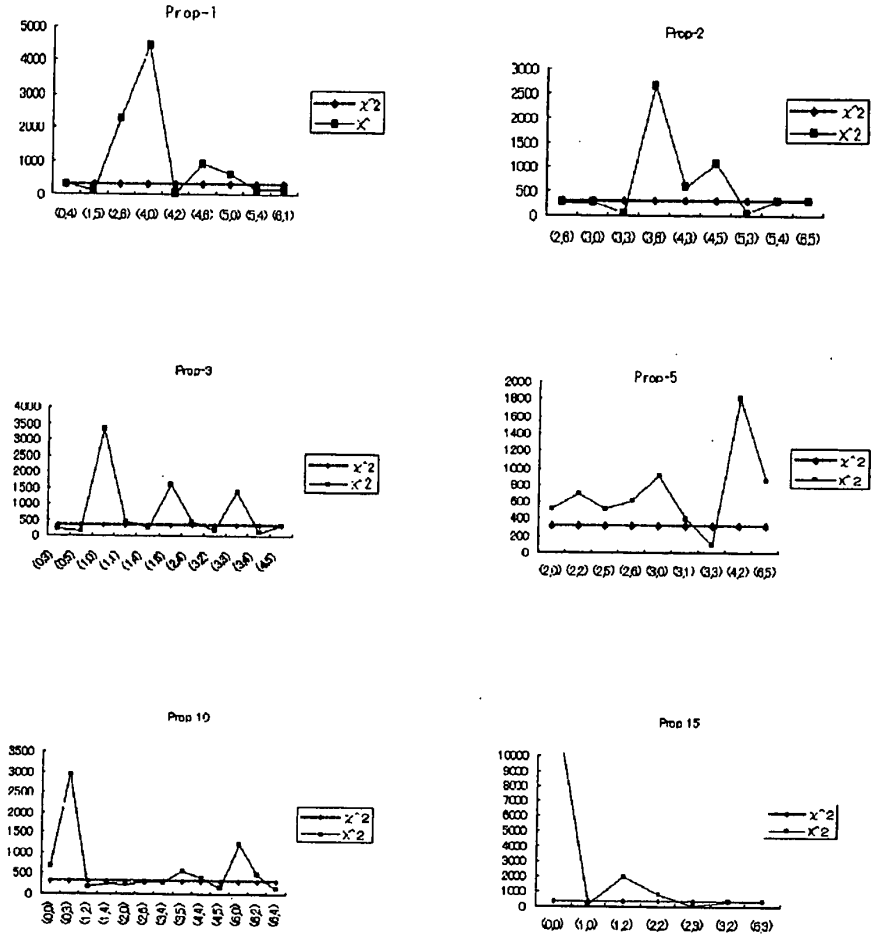


図 6: $\mathcal{X}^2$ test of LVQ(1), LVQ(2); LVQ(3), LVQ(5); LVQ(10) , LVQ(15)