東北薬科大学研究誌, **50**, 41 − 61 (2003)
Journal of Tohoku Pharmaceutical University, **50**, 41 − 61 (2003)

ニューラルネットワークによるアプローチ
ー構造活性相関および分類問題の予測精度改善などを目指して

佐藤　憲一

Neural Network Approach to improve the Prediction Ability
― Quantitative Structure-Activity Relationships and Classification Problems

Kenichi SATO

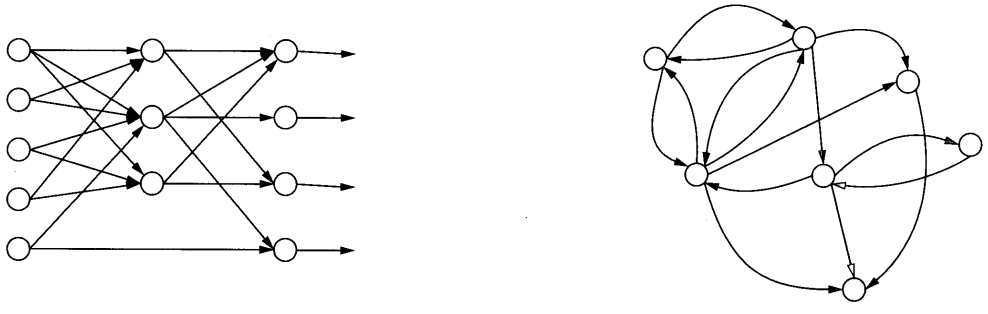
(Received November 22, 2003)

1．はじめに

薬物の分類や生物活性とその分子構造との相関（QSAR）を調べるため70～80年代に広く用いられてきた統計学的手法は多重線形回帰法（MLR）と部分最小2乗法（PLS）のような線形回帰の手法である．[1-3] これらの問題は与えられた対象をデータの特徴をもとに分類するよく見られる作業であり，パターン認識と総称される問題の範疇に入る．これらの方法はそれなりの成果をあげているが，回帰が不良設定問題であることに由来する不安定性や，上記のような薬学分野で取り扱われる問題には構造と分類の関係には強い非線形性がある場合も多いため十分ではなかった．[4] また，臨床検査値などのデータから疾病を分類する診断問題などの場合も分布に大きな偏りの見られる場合が多いため，多重線形回帰などの従来の統計的手法が実用に耐える十分な分類方法となることは困難であった．[5]

90年代に入ると脳の神経回路をモデル化したニューラルネットワークが医療・薬学分野にも導入され，線形，非線形，大量データ，などに対応できるこの手法がコンピュータ中に実装され，コンピュータの計算・記憶能力の著しい発展とも相まって，構造活性相関や分類問題のより進んだ解析が可能になり，現在，その発展は加速している．

ニューラルネットワークの登場は，それまでの主流であった統計学の多くの手法が多変量解析というガウス分布を暗黙のうちに想定する枠の中にあり現実の問題への適用面には制約が厳しいためより一般的なパターン認識の手法が望まれていたことから，どういう仕組みで識別のアルゴリズムを構築したらよいかわからない多くの問題に光明をもたらすものであった．[6] パターン認識は仕組みが複雑で一筋縄でゆくものではなく，どの手法がよいかは問題によって異なる．従って，ニューラルネットワークの手法を医療・薬学分野の問題に適用する場合も，さまざまな条件のもとでの個々の問題を多様な視点で分析してみる必要がある．それらの積み重ね



Multi-layer neural networks　　Mutual connection neural networks

Fig. 1.　Types of neural network

を通して，個々の問題のより適切な処理や全体の統一的な認識も可能となるものと期待できる．

ニューラルネットワークの基本構成としては階層型と相互結合型が代表的である[7]が，これまでに医療・薬学分野の分類問題やQSAR解析に応用されているのは多くは階層型であり，特に逆伝播型（Back-Propagation）学習に基づくニューラルネットワーク（BPNN，以下BPと略）がその中心であった．[4] 本稿では，BPを話題のベースとしてニューラルネットワークの基本と応用について概説し，さらに，その発展型であるベイジアン正則型ニューラルネットワーク（BRNN），データ分類の可視化能力にすぐれた自己組織型ニューラルネットワーク（SONN）の医療・薬学分野の問題への応用について，最近の発展に焦点をあてて解説する．

2．階層型ニューラルネットワークによる解析

2．1 階層型ニューラルネットワークとBP

ニューラルネットワークは多数のニューロン（簡略化したモデルニューロンは，ノードとも呼ばれる）をシナプス結合させ，入力情報に応じてノード間の結合の強さ（結合荷重）を調整することにより情報処理を実現する．

Fig. 2は医療・薬学系のQSARや分類問題などによく利用される3層のネットワークで，ある薬物の構造記述変数（m個）を入力データとして，それぞれ入力層のm個のノードに入力する．この信号はノード間結合を通して中間層のノードに伝播される．中間層の各々のノードから出力される信号もノード間結合を通して出力層のノードに伝わる．出力層には分類するクラス数だけのノードを用意しておけば，どのノードが信号を出力するかを見てそのデータが属するクラスを知ることができる．1つのノードに信号$\mathbf{x}$が入力したときの出力を$f(\mathbf{x})$とすると，多数のシナプス結合（結合荷重w_i）をもつノードへの入力は，シナプス前ノードiからの出力をx_iとすると，$\mathbf{w}\cdot\mathbf{x} = \sum w_i x_i$なので，このノードの出力は$f(\mathbf{w}\cdot\mathbf{x})$である．結局，3層のネットワークへ信号$\mathbf{x}$が入力したときの出力層のあるノードの出力$y$は，層1→層2の結合荷重を$w_1$かつ層2→層3の結合荷重を$w_2$として，次式で与えられる．

$$y = f(w_2 f(w_1 \cdot \mathbf{x})) \qquad (2\text{-}1)$$

ノードの動作関数$f(\ )$の選び方が重要で，もし$f(x) = x$と設定すれば入力$\mathbf{x}$と出力yは線形の関係に従う．その場合，入力層と出力層の2層のみからなるネットワークはMLRと同じ振る舞いを示すことが知られている．[4] 一般には，

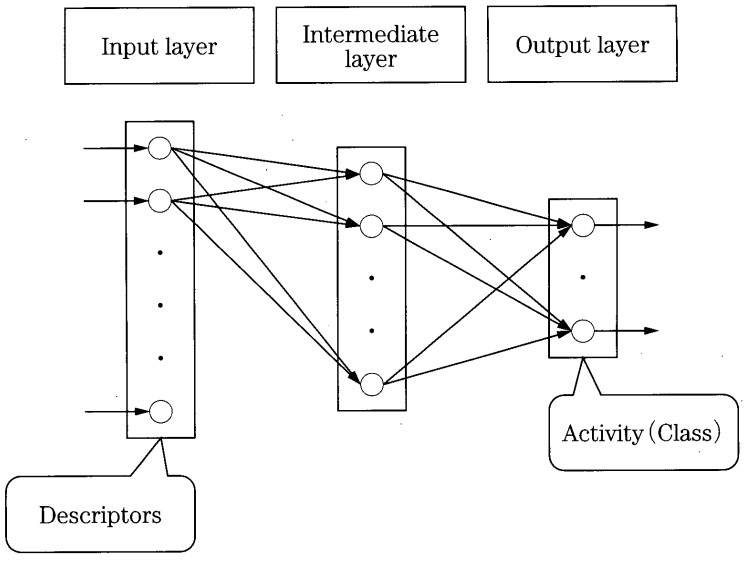


Fig. 2. Multi-layer feedforward neural network

シグモイド関数 f(x) = 1/(1+exp(−αx)) を使用することが多く，この場合，式（2-1）の入力 x と出力 y の関係は非線形となる．f としてはガウス関数や非単調関数なども用いられる．[7,8] 2 層からなるネットワークでも動作関数にシグモイド関数を用いる場合は MLR より優れた結果が得られる．

このようにネットワークの入出力関係を決定するものは，ニューロンの動作関数の形とシナプス結合荷重の大きさである．動作関数をシグモイドなどに決めれば，ニューラルネットワークの振る舞いを決めるのは結合荷重 w ということになる．80 年代後半に Rumelhart らが誤差逆伝搬学習法という結合荷重 w を決定するための実際的なアルゴリズムを提案してから[9]，実用的な手法としてのニューラルネットワークの広範な応用が始まったのである．今，データ D = $\{\mathbf{x}^{(p)}, t^{(p)}\ (p=1 \sim N)\}$（x は構造記述変数，t は生物活性値，N は学習に用いられる誘導体の数）が与えられた時，ネットワークの出力 y と教師信号 t から作られる誤差関数

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_p (t^{(p)} - y(\mathbf{x}^p;\mathbf{w}))^2 \qquad (2\text{-}2)$$

を最小にする w がネットワークの振る舞いを決める最適なパラメータとなる．これは誤差最小化原理と呼ばれ，ネットワークの出力 y が教師信号 t に近づくように結合荷重 w を修正する式を導く．

2．2　カルボキノン誘導体の QSAR

カルボキノン誘導体は，Chart 1 に示す構造をもつ化合物群である．それらは mitomycin C の構造をもとに中尾らによって合成された物質で一般に抗ガン活性を有し，一部のものは実際に医療に用いられている．[10] Hansch 法を用いた古典的な QSAR 解析では次式のように生物活性の強さと薬物の構造変数の間に線形関係を仮定する．[1,2]

$$\log(1/C) = a\times MR_{1,2} + b\times\pi_{1,2} + c\times\pi_2 + d\times MR_1 + e\times F + f\times R \qquad (2\text{-}3)$$

ここで log(1/C) は生物活性値であり，投薬によって 40% 延命効果を示す薬量（MED）のマウス 1 kg 当たりのモル濃度を C とした．[11] また，分子屈折定数（MR），疎水性定数（π），置換基定数（F, R），立体効果と全疎水性定数を推定するための $MR_{1,2}$，$\pi_{1,2}$ は 6 種の物理化学パラメータである．係数 a, b, c, d, e, f は回帰式から決められる．

この QSAR の問題に階層型ニューラルネットワークを適用する場合は，(2-3) のような関係式は仮定せず，直接，構造記述変数と生物活性の関係をネットワークに学習させればよい．学習が終了すれば最適なネットワークの結合荷重 w が確立するので，ネットワークに未知化合物の構造変数を与えて，式 (2-1) から活性値を予測させることができる．入力層，中間層，出力層にそれぞれ，7, 12, 1 個のノードを用意して，逆伝播型の学習が終了後，ニューラルネットワークはほとんど完全な fitting が可能であり，多重回帰法よりすぐれていた．[12]

QSAR 解析や薬学系の分類問題ではサンプル数が 10 組～ 100 組程度と少ないことも多い．その場合には計算パワーさえあれば，学習サンプルのみから比較的簡単に予測性能を評価できる方法に resampling 手法があり，その中でも最も単純な手法である leave-one-out 法がよく利用される．Leave-one-out 法では N 個のサンプルが与えられた場合，それを $N-1$ 個の学習サンプルと 1 個のテスト用サンプルとに分割し，$N-1$ 個の学習サンプルを用いた学習結果で 1 個のテスト用サンプルを評価する．このような分割

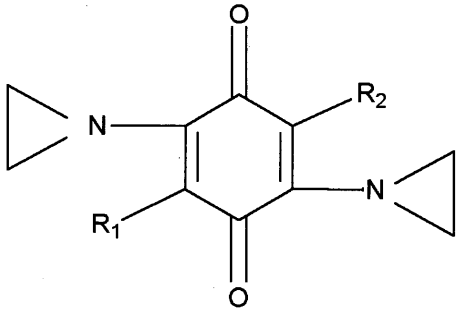


Chart 1.　Structures of Carboquinones

Table 1. Physico-chemical parameters, Observed Biological Activities for 37 Caroboquinone Derivatives

	$MR_{1,2}$	$\pi_{1,2}$	π_2	MR_1	F	R	obs
1	5.08	3.92	1.96	2.54	0.16	-0.16	4.33
2	4.5	3.66	3.16	0.57	-0.08	-0.26	4.47
3	4.86	5	2.5	2.43	-0.08	-0.26	4.63
4	3	2.6	1.3	1.5	-0.08	-0.26	4.77
5	3.57	2.51	2.01	0.57	-0.12	-0.14	4.85
6	3	3	1.5	1.5	-0.08	-0.26	4.92
7	3.79	2.16	1.66	0.57	-0.08	-0.26	5.15
8	6.14	0.72	0.36	3.07	-0.08	-0.26	5.16
9	2.06	2	1	1.03	-0.08	-0.26	5.46
10	2.28	1.03	0.53	0.57	-0.08	-0.26	5.57
11	1.58	-0.04	-0.02	0.79	0.52	-1.02	5.59
12	2.07	1.8	1.3	0.57	-0.08	-0.26	5.6
13	4.24	0.98	-0.52	1.5	-0.04	-0.13	5.63
14	1.14	1	0.5	0.57	-0.08	-0.26	5.66
15	1.6	1.3	1.3	0.1	-0.04	-0.13	5.68
16	2.75	1.53	1.03	0.57	-0.04	-0.13	5.68
17	3.56	1.45	-0.05	1.5	-0.08	-0.26	5.68
18	3.42	1.03	0.53	1.71	-0.08	-0.26	5.69
19	4.23	0.98	-0.02	1.03	-0.04	-0.13	5.76
20	2.78	1.23	0.73	0.57	-0.08	-0.26	5.78
21	1.96	2	1.5	0.57	-0.08	-0.26	5.82
22	1.6	1.5	1	0.57	-0.08	-0.26	5.86
23	4.45	0.01	-0.49	0.57	-0.04	-0.13	6.03
24	3.09	0.75	0.25	0.57	-0.08	-0.26	6.14
25	3.77	0.48	-0.52	1.03	-0.04	-0.13	6.16
26	3.55	1.25	0.75	0.57	-0.08	-0.26	6.18
27	3.77	0.48	-0.02	0.57	-0.04	-0.13	6.18
28	3.09	0.95	0.45	0.57	-0.08	-0.26	6.18
29	2.63	0.45	-0.05	0.57	-0.08	-0.26	6.21
30	3.09	0.95	-0.05	1.03	-0.08	-0.26	6.25
31	1.78	0.34	-0.16	0.57	-0.08	-0.26	6.39
32	3.09	0.75	0.25	0.57	-0.08	-0.26	6.41
33	3.31	-0.02	-0.52	0.57	-0.04	-0.13	6.41
34	1.66	0.18	0.18	0.1	0.1	-0.92	6.45
35	2.42	-0.32	-0.16	1.21	-0.08	-0.26	6.54
36	2.13	0.68	0.18	0.57	0.06	-1.05	6.77
37	2.47	-0.13	-0.63	0.57	-0.04	-0.13	6.9

の仕方はN通りあるので，その全てに対するテスト結果の平均を計算し，それを予測性能の評価値として利用する．また，学習によって得られた結合荷重 w が，学習していない問題に対してどの程度一般性をもっているか，つまり，学習していないデータをうまく処理できる能力を汎化能力と呼んでいる．

Leave-one-out 法による汎化性能のチェックでも，Fig. 3 に示すように高い予測能力を示している．図で白丸印は leave-one-out 法により予測された値である．直線は観測値であり，直線より上位の値は観測値より強く，また下位は弱く予測されたものである．[12)]

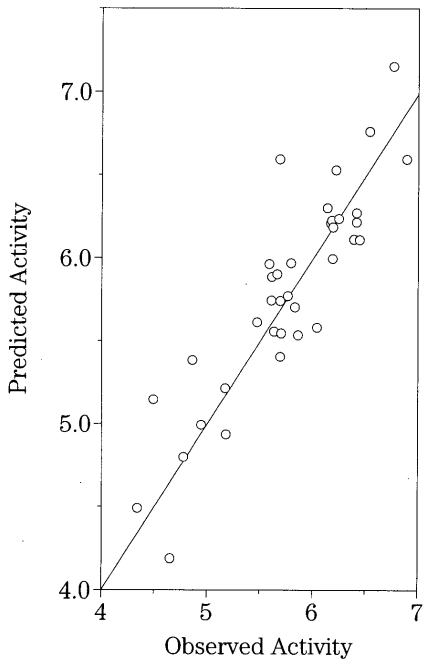


Fig. 3. Predicted activity versus Observed activity for Carboquinones

2．3　BPネットワークの抱える問題

誤差関数の最小化原理に基づくBPは，複雑な問題に対処する際に速い，シンプルかつ柔軟であるというすばらしい長所を持っているが，同時に，次のようなさまざまな問題も内包している．[13]

(1)誤差最小化，もっと一般的には最尤推定法という決定論に基づき結合荷重wを決める場合，誤差関数$E_D(w)$は一般にはw空間での複雑な関数でローカルな極小点も多いため，実際に$E_D(w)$ →ゼロを満たすwの最適値を求めるのは困難なデータが一般的である．また，E(w) →ゼロの手順をコンピュータ上に実装して，効率よくwの最適値を探すためにはさまざまな数値計算上の工夫も必要である．

(2)非線形の問題を扱えるという強力な長所の裏返しとして，overfittingにより汎化能力の低下を招きやすい．

(3)中間層のノード数や結合荷重w以外のネットワークに現れるパラメータ値などはさまざまな組み合わせが可能であり，最適値を決める（つまり，モデル選択）のは容易でない．

(4)式（2-1）の入力xと出力yの関係はこのままではブラックボックスである．単なる予測だけでなく，予測のメカニズムを知るためにはwの内部構造を知る必要があり，そのためにはwはシンプルなのが望ましい．

これらの問題にも関連して，ρ＝（サンプルデータの数/結合荷重の数）という比を考え，

$$\text{「rule of thumb」}\quad \min \leqq \rho \leqq \max \qquad (2\text{-}4)$$

という条件がBPの適用成功には必要という経験則がよく指摘される[14-17]が，対象となるデータによりmin，maxの値は異なり，バラついている．サンプルデータが少ないのに結合荷重の数が多いと適切なwは決まらないだろうし，サンプルデータが多いのに結合荷重の数が少ないと非線形に由来するoverfittingが生じて汎化能力が低下するだろうから，この条件はある程度妥当であろうが，現時点では定量的な条件にはなり得ず，おおまかな指針を与える程度に過ぎない．

2．4　ベンゾジアゼピンのQSAR

ベンゾジアゼピンはChart 2に示す構造をもち抗不安薬として広く用いられている．ベンゾジアゼピンの作用は，脳におけるGABA作動性伝達の増強によるもので，ベンゾジアゼピンがその受容体に結合すると，GABAとその受容体間の結合親和性が高まることが知られている．ベンゾジアゼピンはさまざまな手法で広くQSAR

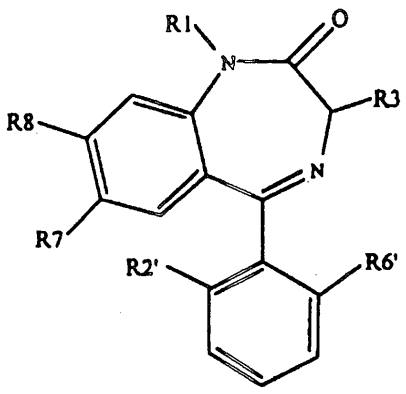


Chart 2. Structures of Benzodiazepine

Table 2. Benzodiazepine 57 derivatives

Name	R_7	R_1	$R_{2'}$	$R_{6'}$	$R_{3'}$	R_8	log IC_{50}[a]
Ro 05-3061	F	H	H	H	H	H	1.602
Ro 05-4865	F	Me	H	H	H	H	1.230
Ro 05-6820	F	H	F	H	H	H	0.869
Ro 05-6822	F	Me	F	H	H	H	0.708
nordazepam	Cl	H	H	H	H	H	0.973
diazepam	Cl	Me	H	H	H	H	0.908
Ro 05-3367	Cl	H	F	H	H	H	0.301
delorazepam	Cl	H	Cl	H	H	H	0.255
Ro 07-9957	I	Me	F	H	H	H	0.462
Ro 05-2904	CF_3	H	H	H	H	H	1.114
Ro 14-3074	N_3	H	F	H	H	H	0.724
nitrazepam	NO_2	H	H	H	H	H	1.000
Ro 05-4435	NO_2	H	F	H	H	H	0.176
flunitrazepam	NO_2	Me	F	H	H	H	0.580
clonazepam	NO_2	H	Cl	H	H	H	0.255
Ro 05-4082	NO_2	Me	Cl	H	H	H	0.342
Ro 05-5390	NO_2	H	CF_3	H	H	H	0.544
Ro 20-7736	NHOH	Me	F	H	H	H	1.982
Ro 05-3072	NH_2	H	H	H	H	H	2.587
Ro 05-3418	NH_2	Me	H	H	H	H	2.663
Ro 20-1815	NH_2	Me	F	H	H	H	1.813
Ro 05-4619	NH_2	H	Cl	H	H	H	1.875
Ro 05-4528	CN	Me	H	H	H	H	2.580
Ro 20-2541	CN	Me	F	H	H	H	1.477
Ro 20-2533	Et	H	H	H	H	H	1.556
Ro 20-5747	$CH=CH_2$	H	H	H	H	H	1.380
Ro 20-5397	CHO	H	H	H	H	H	1.633
Ro 20-3053	COMe	H	F	H	H	H	1.255
Ro 05-2921	H	H	H	H	H	H	2.544
Ro 05-4336	H	H	F	H	H	H	1.322
Ro 05-4520	H	Me	F	H	H	H	1.146
Ro 05-4608	H	Me	Cl	H	H	H	0.580
halazepam	Cl	CH_2CF_3	H	H	H	H	1.964
Ro 06-9098	NO_2	CH_2OCH_3	H	H	H	H	2.633
Ro 20-1310	Cl	$C(CH_3)_3$	H	H	H	H	2.792
Ro 07-2750	Cl	$(CH_2)_2OH$	F	H	H	H	1.389
Ro 22-4683	NO_2	$C(CH_3)_3$	F	H	H	H	2.477
Ro 07-4419	H	H	F	F	H	H	1.279
Ro 07-3953	Cl	H	F	F	H	H	0.204
Ro 07-4065	Cl	Me	F	F	H	H	0.613
Ro 07-5193	Cl	H	Cl	F	H	H	0.477
Ro 22-3294	Cl	H	Cl	Cl	H	H	0.845
Ro 07-5220	Cl	Me	Cl	Cl	H	H	0.740
Ro 13-3780	Br	Me	F	F	H	H	0.380
Ro 11-4878	Cl	H	F	H	Me	H	0.544
meclonazepam	NO_2	H	Cl	H	Me	H	0.079
Ro 11-6896	NO_2	Me	F	H	Me	H	0.845
Ro 06-7263	Cl	Cl	H	H	Me	H	1.690
ozazepam	Cl	H	H	H	OH	H	1.255
temazepam	Cl	Me	H	H	OH	H	1.204
lorazepam	Cl	H	Cl	H	OH	H	0.544
Ro 20-7078	Cl	H	F	H	Cl	H	0.724
Ro 07-6198	H	H	F	F	H	Cl	1.447
Ro 20-8895	H	H	F	H	H	Me	1.279
Ro 22-6762	Cl	Me	H	H	H	Cl	1.602
Ro 20-8065	Cl	H	F	H	H	Cl	0.556
Ro 20-8552	Me	H	F	H	H	Cl	1.146

[a] IC_{50} in nmol L^{-1}.

解析がなされている化合物であり，BP による QSAR 解析もいくつか行なわれている．[14-16]

Table 2 に挙げた 57 個のベンゾジアゼピン誘導体のセットは 6 個の置換基 R_1, $R_{2'}$, R_3, $R_{6'}$, R_7, R_8 の位置にさまざまな官能基を配置したもので，各置換基を特徴づける物理化学パラメータとして，分子屈折定数（MR），疎水性定数（π），極性定数（F），共鳴定数（R），芳香属双極子（μ），Hammet メタ定数（σ_m），Hammet パラ定数（σ_p）の 7 つがよく用いられる．分子の化学構造やその他の特性を適切に表現し定量的な情報とする方法はいろいろ試みられているが，ここでの置換基の構造を種々の物理化学パラメータで置き換えるやり方は，分子の性質は置換基などの部分構造の性質の総和で決まるという基本的な考え方による．薬物の受容体や酵素との結合は，薬物の大きさならびに電子密度などの電子的効果が重要な役割を果たし，また，その生物活性発現は，これらの効果の総合的結果と考え，それは，分子を構成する部分構造の疎水性のパラメータ，置換基の大きさのパラメータ，電子的効果の各パラメータの寄与の総和で説明できると仮定するのである．物理化学パラメータの値は，測定値が用いられることが多く，この場合は利用できるものが限られている．量子化学の多体問題として，コンピュータを用いた

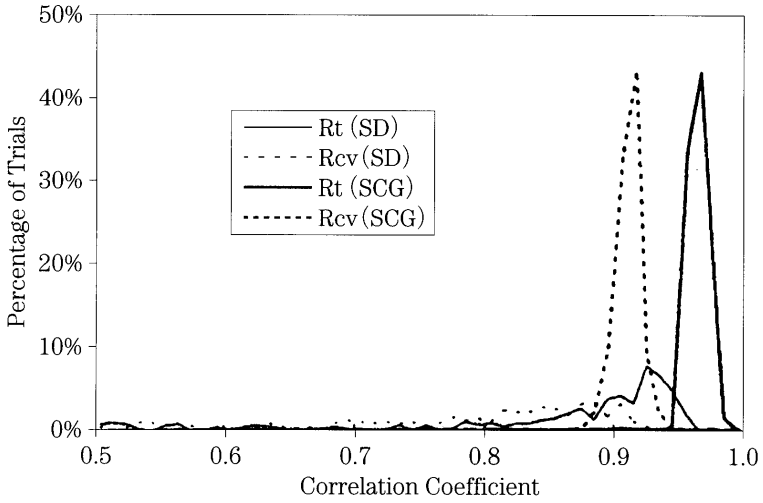


Fig. 4. Percentage of trials versus correlation coefficient

計算により，原子間相互作用から直接的あるいは半経験的に分子を特徴づける諸量や物理化学パラメータを求めることもできるが，精度の高いものを容易に求めることが課題である．

誤差最小化原理を具体的にコンピュータ上に実装する場合には，さまざまな実装手法を真の最適値からのずれの大小，要する計算時間の大小，安定性などの，現実的な側面から十分検討する必要がある．BP のプログラミングにはその簡潔さから，最急降下法（Steepest Descent）がよく用いられるが，結合荷重 w のさまざまな初期値から出発して（誤差が基準値以内という意味で）収束した w には真の値からのバラつきがある．共役勾配法（Scaled Conjugate Gradient）ならより良い w が安定して得られることもわかっている．

Fig.4 はベンゾジアゼピンでの両手法に対する予測精度の分布を比較したものである[15]が，この事実を明快に示している．この他にも誤差関数の最小化を実現するための多くの改良がなされており，例えば筆者らは，結合荷重 w のみならず動作関数に含まれるパラメータも可変なものとして評価関数を最小化することや，動作関数に非単調なものを使用することで計算時間が大幅に短縮できる場合があることを見出している．[8]

ニューラルネットワークは学習データを通して入力（構造変数）と出力（活性値や分類クラス名）の関係を学び，未知のデータに対する入力が与えられた時の，正しい出力を予測する．ニューラルネットワークを構成する結合荷重 w_{ij} には大きいものも小さいものもあり，かなり小さい w_{ij} は一般には重要でないと考えられる．この考えを推し進めて，学習過程中にかなり小さい w_{ij} を積極的に減じる操作を行ない，結果としてシンプルな繋がりを持つネットワークを構成することを“刈り取り”（pruning）と呼ぶ．これまでの BP に関する多くの研究から，適切な pruning はネットワークの複雑さを減らしその汎化性能を高めるのに大変有効であることがわかっており，再構築学習をはじめ，小さい w_{ij} を積極的に減じる操作法が提案されている．[18-20] Pruning は入力から出力に至る情報の主要な流れを明らかにする役割を持つので，活性値や分類クラス名を決めるのにどの構造記述変数の影響が大きいかという，ネットワークから予測された結果の解釈，従って，活性の大きい新薬の構造を予測するなどのためにも大切である．特に，入力層のノードで結合荷重を削除するものがある場合，入力信号である構造変数のセットから一部のものを不要と見なすことになるので変数選択と呼ばれる．Maddalena ら[14]はベンゾジアゼピンの 6 個の置換基それぞれに付随する 7 個の物理化学パラメータ MR, π, F, R, μ, σ_m, σ_p の総計 42 個から出発して，変数を順次 1 ずつ減じた計算を行ない，残すべき重要な変数として 10 個の物理化学パラメータを選び出した．

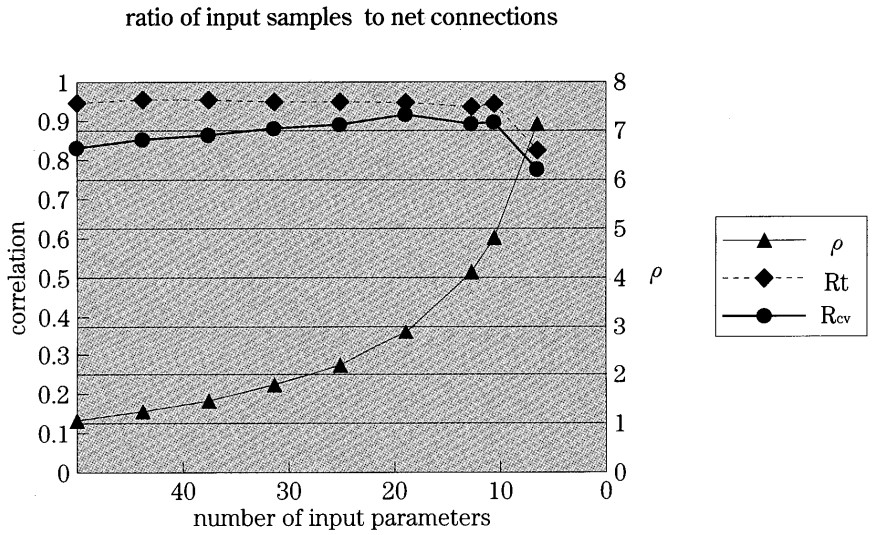


Fig. 5. Cross-validated correlation coefficient (R_{cv}) and ratio of input samples to net connections (ρ) as a function of the number of descriptors.

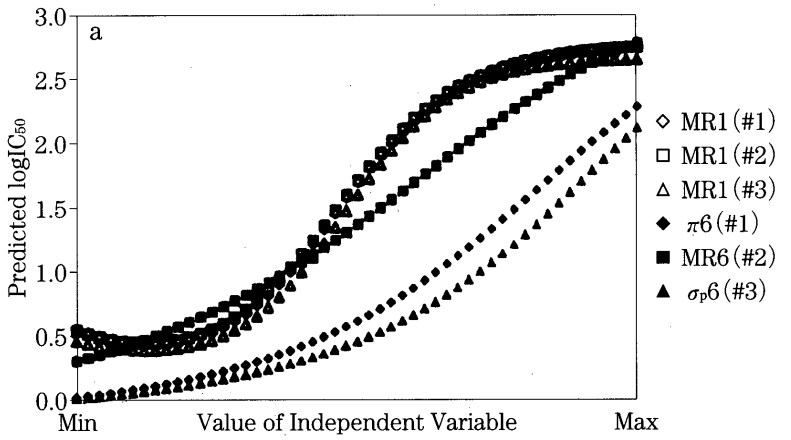


Fig. 6. Predicted activity as a function of the descriptors that have been chosen by GNN. The low value in vertical axis corresponds to high activity.

Fig. 5 には選択した物理化学パラメータ数と予測された活性値の精度（観測値と予測値の相関係数 R_{cv}）の関係が leave-one-out 法により示されており，パラメータ数 10 ～ 15 個，$3 \leqq \rho \leqq 5$ において汎化性能が最も優れていることがわかる．また，この計算では中間層に 4 個のノードを配置した時の結果がベストであった．中間層のニューロンを十分多くとれば任意の関数を表現できることが舟橋らにより数学的には証明されているものの，中間層のニューロン数が多くなると overfitting となりネットワークの汎化性が低くなる傾向がある．So らは[15,21]変数選択に遺伝的アルゴリズムを用いることでさらに優れた汎化性能が得られることを示し，最重要な 6 個の変数（7 位の π，σ_m，1 位の MR，2' 位の σ_m，σ_p, 6'位の σ_p）を選び出している．

Pruning により重要な変数が選び出されると，その中の 1 つの物理化学パラメータを変化させて（残りは活性値の大きい化合物と同じ官能基に固定する）Fig. 6 のように高い活性を与える

パラメータ値がわかれば，さまざまな官能基とその物理化学パラメータ値を載せたライブラリから，受容体との親和性を増加させる官能基を見つけることが可能である．このように，ニューラルネットワークはすぐれたQSAR解析の手段となり，薬物－受容体結合に親和性の高い構造が持つ物理化学パラメータ値への洞察を与え，そこから鍵となる構成要素＝官能基を示唆することができる．Soらは7位の官能基を CH_2CF_3, SO_2F, 2'位の官能基を NO_2, SO_2F などに置き換えることで優れたベンゾジアゼピン誘導体を予測した．

化学構造を表現するのに物理化学パラメータを使用する以外に，Molecular Index（分子指標）もよく使用される．これは化学構造式や立体化学に着目して構成原子の原子価や原子間結合数などを用いて分子の種々の特性を数値化して表現するもので，Randicによる結合指標の考案以来，Kier and Hall Index, Atomistic Indexをはじめとして分子構造の表現精度を高めることを目指して多数の拡張がなされている．[16,22] また，官能基としてそのまま表現することが必須であるものも知られている．ベンゾジアゼピンのBPによるQSAR解析は構造変数に分子指標を用いた計算も，最近，行なわれており，R, K, Aなどの分子指標をセットで用いると物理化学パラメータを用いた場合と同程度に良い結果が得られた．[16]

3．ベイズ推論を取り入れた正則化ニューラルネットワーク

3．1　ベイジアンニューラルネットワーク

薬物の分類やQSAR研究においてBPは，従来の統計手法の困難を乗り越えて，大きな役割を果たしている．医療・薬学系の複雑な問題に対処する際に速い，シンプル，柔軟という長所は魅力的である．しかし，現実に存在する巨大な，偏りのある，ノイズを含む，……といったデータを処理するのにはBPではやや力不足な面もあり，2節で述べた問題点(1)～(4)がある．決定論的な学習アルゴリズムで複雑なデータを処理する場合には極小値へトラップされなかなか最適解が得られないという困難は大きな問題であり，ネットワークのサイズが増大するにつれてより深刻になることが，たとえば，「何千種にもおよぶ漢字の認識」といった問題ではよく知られており[7]，医療・薬学系の多くの問題でも避けられない可能性が高い．

これらの問題を克服するため，BP型ニューラルネットワークにベイズ統計の枠組みを取り入れて拡張したニューラルネットワークがMackayにより提案され[23]，ベイジアンニューラルネットワーク（Bayesian Reguralized Neural Network, BRNN）と呼ばれている．回帰問題をベイズ流に扱うには誤差関数 E_D と正則化項 E_W からそれぞれ作られる尤度と事前分布

$$P(D|\vec{\omega},\beta)=\frac{1}{Z_D(\beta)}\exp(-\beta E_D),$$

$$P(\vec{\omega}|\alpha)=\frac{1}{Z_W(\beta)}\exp(-\alpha E_W) \qquad (3\text{-}1)$$

からベイズの定理により事後分布は

$$P(\vec{\omega}|D,\alpha,\beta)=\frac{P(D|\vec{\omega},\beta)\,P(\vec{\omega}|\alpha)}{P(D|\alpha,\beta)}$$

$$=\frac{1}{Z_M}\exp(-M(\vec{\omega})) \qquad (3\text{-}2)$$

となる．ここで $M(\omega)=\beta E_D+\alpha E_W$ は目的関数であり，α は正則化，β^{-1} はノイズの，それぞれ強さを表わすハイパーパラメータである．目的関数に対するガウス近似が許される条件下ではMacKayによりシンプルな解析式が得られている．一般的には事後確率を含む積分の評価が非常に難しいのであるが，Nealによりマルコフ連鎖モンテカルロ法が導入され，より現実的な問題へのアプローチが可能になっている．[24]

BRNNの枠組みでは，構造記述変数の中でより重要な成分を自動的に決定することが可能であり自動関与度決定（Automatic Relevance Determination, ARD）と呼ばれている．[23] それには構造記述変数の数だけの正則化のハイパー

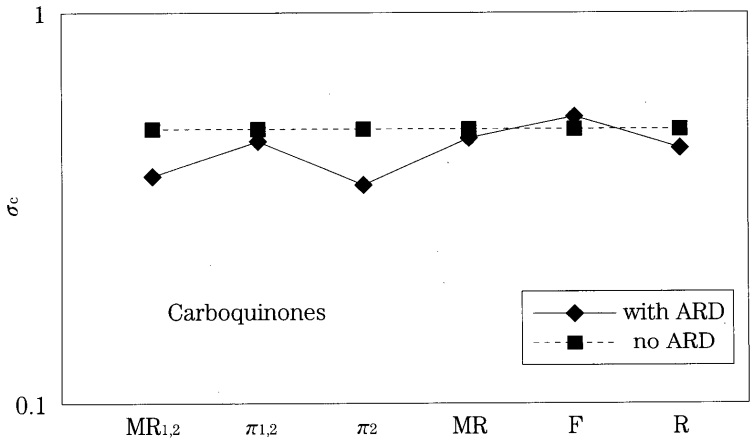


Fig. 7. Values of hyperparameter $\sigma_c = \alpha_c^{-1/2}$ for six physicochemical parameters.

パラメータ α_c(c = 1 ～ n) を導入すれば,

$$\exp(-\alpha E_W) \rightarrow \exp(-\sum_c \alpha_c \sum_{i \in c} \frac{\omega_i^2}{2}) \quad (3\text{-}3)$$

α_c の小さい入力成分ほど，結合荷重 $w_{ij}^{(1,2)}$ が大きいので，重要であることがわかる．ARD は，変数選択法がハードな pruning であるのに対し，ソフトなネットワークの pruning であり，確率理論に基礎を置く点や自動的決定である点は他の pruning に比べて優れている．

新しいデータに関する予測値の分布はすべてのハイパーパラメータをまとめて $\vec{\gamma}$ と記せば

$$P(t^{(N+1)}|D) = \int P(t^{(N+1)}|\vec{\omega}, \alpha\vec{\gamma}) P(\vec{\omega}, \vec{\gamma}|D) d\vec{\omega} d\vec{\gamma} \quad (3\text{-}4)$$

で与えられ，出力 y の予測値，2 乗誤差はそれぞれ次のようにおけばよい．

$$P(t^{(N+1)}|\vec{\omega}, \vec{\gamma}) \rightarrow y(\mathbf{x}; \vec{\omega}), (y(\mathbf{x}; \vec{\omega}) - \bar{y})^2 \quad (3\text{-}5)$$

BRNN では学習により結合荷重 w の最もらしい確立分布 P(w) が与えられるのであり，P(w) が結合荷重 w の最適値近傍以外ではゼロとなる極限では BP の結果に帰着するが，一般には BP のように決定論的に最適値のみが決まるわけではない．マルコフ鎖モンテカルロ法によるサンプリング手法が，決定論的な学習アルゴリズムではよく遭遇するローカルな極小値へのトラップという問題を避け，またサンプリング出力による適合値と予測値へのベイジアンモデルでの平均操作という自然な形でデータ不確実性が考慮される．このことが，正則化項の導入と相まって，ベイジアンニューラルネットワークに頑強性と柔軟性を付与するのである．BRNN の医療・薬学系の分類問題や QSAR への適用例は未だ少ないが，ある程度の有効性は確立されつつあるので最近の進展について，以下に述べる．

3．2 いくつかの QSAR 解析

筆者らは Neal の BRNN を QSAR の問題に応用し，2 節で取り上げたカルボキノン誘導体の解析を行なった．[25] Leave-one-out 法による活性値の予測精度は BP による計算結果とほぼ同等であったが BRNN では計算の安定性が高い．また，ARD により使用した物理化学パラメータの重要度を定量的に評価できるだけでなく予測性能の改善が見られた．これは ARD による pruning が自動的に行われた結果，重要度の低い構造記述変数の影響を排除することにより汎化性が高まることを示唆している．Fig.7 に構造記述変数として使用された物理化学パラメータ $MR_{1,2}$, $\pi_{1,2}$, π_2, MR, F, R に関係する結合荷重に対するハイパーパラメータ $\sigma_c = \alpha_c^{-1/2}$ を示す．

BRNN が「多くの化合物を医薬品ライクと非

医薬品ライクに分類する」ことができるかという問題がそれぞれ数万にも及ぶ非医薬品ライクな分子からなるデータベースと医薬品ライクな分子からなるデータベースを用いて検討された.[26] 構造記述変数として，分子量などの7個の分子全体に関する情報と166個のISISキーによる分子内の特定の官能基に関する情報を用いた．このようにサンプル数も巨大で，かつ構造記述変数もかなり多い分類問題にBPを使うことは困難であるが，NealのBRNNを用いて医薬品ライクな化合物を80% ～ 90%は正確に予測することができ，優れた汎化能力も示された．ARDにより自動的に変数選択も行なわれている．このモデルはコンビナトリアルライブラリを設計するために有用な可能性があり，1セットの10000分子からサイズ100の医薬品ライクなライブラリを生成するコンピュータ実験によると，ランダムにライブラリを生成する場合に比べて，少なくとも3桁ないし4桁の改良が得られている．

構造変数として分子指標を用いたQSAR解析がBRNN計算によりベンゾジアゼピン，ムスカリンとテトラヒメナ毒に対してなされた.[27-29] 使用された分子指標はTable 3に示すR, K, A, B, G, Fの6種類であり，Gは環，Fは官能基を特徴づける指標である．この計算により，BRNNではoverfittingや過学習も少なく，十分な安定

Table 3. Molecular Indices used in the QSAR analysis

Index	Element	Number of Connections	Atom Type
Atomistic			
A1	Mol. Mass		
A2	H	1	H1
A3	C	2	C2(sp)
A4	C	3	C3(sp_2)
A5	C	4	C4(sp_3)
A6	N	1	N1
A7	N	2	N2
A8	N	3	N3
A9	N	4	N4
A10	O	1	O1
A11	O	2	O2
A12	F	1	F1
A13	Si	2	Si2
A14	Si	3	Si3
A15	Si	4	Si4
A16	P	2	P2
A17	P	3	P3
A18	P	4	P4
A19	P	5	P5
A20	S	1	S1
A21	S	2	S2
A22	S	3	S3
A23	S	4	S4
A24	Cl	1	Cl1
A25	Br	1	Br1
A26	I	1	I1
Extended			
B1	C(Ar)		c
B2	N(Ar)		n
B3	O(Ar)		o
B4	S(Ar)		s
B9	H donor		(N)H
B10	H donor		(O)H
B11	H		N = (O)

Rings	No. in ring
G3	3
G4	4
G5	5
G6	6
G7	7
G8	8
Fragments	
F1	H-O-C
F2	H-O-N
F3	C-O-C
F4	N-O-C
F5	N-O-N
F6	C = O
F7	O = C-N
F8	O = C-O
F9	N = O
F10	O = N = O
Randic[15]	
R1	$^0\chi$
R2	$^1\chi$
R3	$^2\chi$
R4	$^3\chi$
R5	$^4\chi$
Kier And	
K1	$^0\chi^v$
K2	$^1\chi^v$
K3	$^2\chi^v$
K4	$^3\chi^v$
K5	$^4\chi^v$

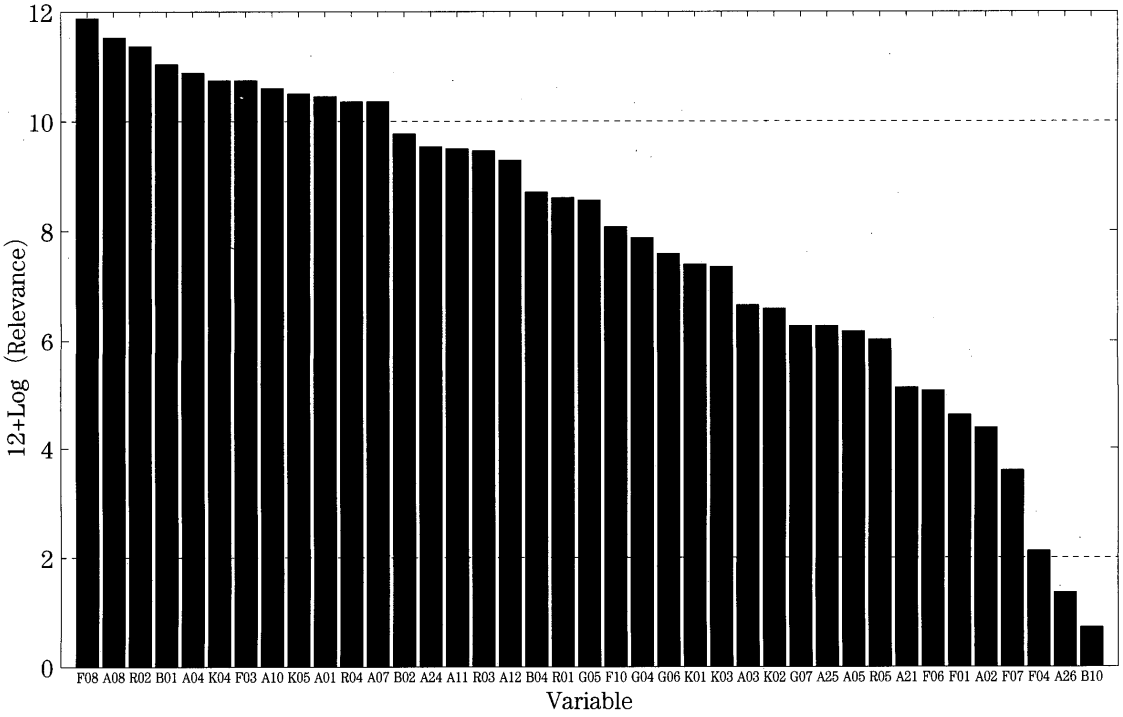


Fig. 8. ARD relevances for the Benzodiazepine derivatives

性と汎化性能をもつことが示された．BRNN は BP に比べてかなりの優位性をもち，また BP と同程度かそれ以上の精度で予測が実現できている．これらの化合物に対しては ARD を適用した計算もなされた．ベイジアンニューラルネットワークでは，ARD の枠組みにより，入力変数の中で雑音となる情報や冗長な情報しか含まないものは正則化項（＝ベイズ統計の事前分布）を考慮する事で除去しているが，これは対応する入力層のニューロンと中間層のニューロンとのシナプス結合を冗長性の程度に応じて減らすことにより，ネットワークの汎化や頑丈さを増している．

ベンゾジアゼピンの場合につき，ARD により得られた分子指標の関与度（relevance）を Fig. 8 に示す．また，ハードな変数選択を併用した MLR 計算の結果によれば，この場合も予測精度は改善するものの，BP や BRNN での pruning による大きな改善と比べれば，わずかである．また，選択の結果重要と見なされた分子指標も両者ではかなり異なっている．これらの事実は，ベンゾジアゼピンの構造記述変数と活性の相関には強い非線形性があることを示唆している．結局，いくら最良の変数選択を実現してみても，解析するデータの内包する非線形性が大きいと MLR のような線形手法では当然ながら解析はうまくいかない．

3．3　臨床医学への応用例

筆者らが最近行なった甲状腺症患者分類の BRNN による解析について述べる．インターネットで公開されている Proben1 内の thyroid データは甲状腺機能障害かどうかの診断が必要な，ある病院を訪れた 1985 年の 3772 人と 1986 年の 3428 人の受検者データであり，受検者のレコードは TSH, T3, TT4, T4U, FT1 の 5 個の臨床検査値に年齢，性別，手術歴，ヨード治療の有無，亢進の外見的兆候，なども数値化して加えた 21 個の項目よりなる．このようなサンプルデータが多い場合は汎化能力の確認に leave-one-out 法などの再サンプリングを行う必要はない．1985 年分のデータをネットワークの学習用に用いて，次年度については各患者のレコード（21 項目）をネットワークに与えてその患者が正常，

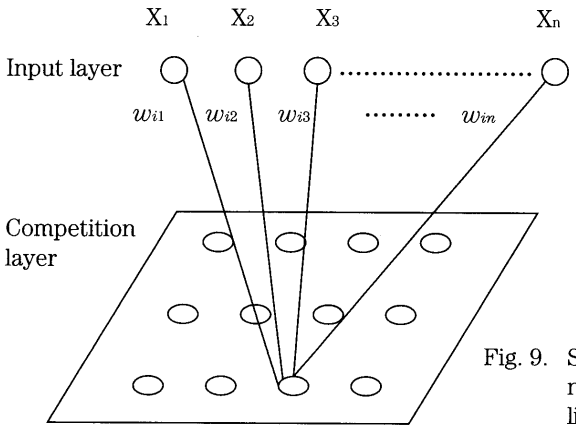
亢進症，低下症のどのケースに該当するかを予測させることができる．このデータは正常が92.6%，亢進症2.3%，低下症5.1%と偏りの大きいデータであるため，ロジスティック回帰分析などの古典的統計手法では異常者の予測が極めて困難であることが知られている．ニューラルネットワークを用いると，BPによる計算でもある程度の予測が可能であることがわかっていたが[5]，BRNN計算を行なった所かなり高い予測正当率が得られ，BPを用いた場合より予測正答率が向上している．また，ARDにより21個の項目中からより判断に重要な役割を果たす項目として，TSH, TT4などが選び出されたが，これは人間による診断において重要視される少数項目に比較的近いと考えられるものである．これらの結果はベイジアンニューラルネットワークが甲状腺機能異常の受検者の分類に有効な手段となることを示唆するもので，他の病気も含めて，BRNNが臨床診断へ応用できる可能性を期待させる．最近，乳癌手術後の予後診断へのBRNNの応用例なども報告されている．[30]

4．特徴のビジュアル化にすぐれた安定なニューラルネットワーク

4．1　SOMとは

階層型のニューラルネットワークとして教師データなしの学習アルゴリズムに従うものも可能で，Kohonenの自己組織化ネットワーク(SONN)[31]がその代表的なものとして知られており，このネットワークではノード同士がHebb則に従って競合学習を行う．Kohonenは感覚器官からの情報にもとづいて脳の皮質に形成されるマッピング形成のメカニズムを単純化したアルゴリズムを開発し，任意の高次元空間におけるデータの低次元空間（特に，2次元の場合が有用なことが多い）への自己組織化地図（Self-Organizing Map，SOM）が生成できることを示した．脳皮質に形成される神経細胞群のマップやKohonenのSOMの本質的な特徴は，高次元データからそれらのもつ最重要な情報を保存する形で解析の容易な2次元マップを作れるということである．そのため，SOMは高次元空間内の複雑なデータ構造がもつ特徴を2次元マップとして視覚的に把握することが威力を発揮する広範な分野の問題に適用され，内在するデータ構造の分析にたいへん有用である．統計学的な視点からみたこのネットワークは，非線形主成分分析の1つであって，非常に多くのパラメータ（ノード間の結合荷重）を使うことで近似的にノンパラメトリックで非線形な主成分分析を実現するものと考えてよい．また，従来のクラスタリング手法の発展したものと見ることも可能である．

医療・薬学分野においても，反応生成物や医薬品の分類問題，分子表面の静電ポテンシャル分布から分子表面の特徴を抽出する問題，[32,33]などへの効果的な応用が報告されているが，他分野での展開に見られるこのネットワークのもつ大きな可能性を考えると，今後の医療・薬学系のさまざまな問題への適用が期待される．SONNのすぐれた点は特徴地図というビジュアルな情報を通して容易に分析する手段を提供すること



$$\vec{\omega}_i(t+1) = \vec{\omega}_i(t) + \alpha(t)\gamma(t)[\vec{\chi}(t) - \vec{\omega}_i(t)]$$

Fig. 9. Schematic representation of Kohonen Self-Organizing network with n input variables. Circles are neurons and each line depicts the connection weights between two neurons.

であり，BPと比べて解法に安定性がある．

Kohonenのニューラルネットワークは Fig. 9に示す2層のネットワークである．[31,34] 第1層はn次元の入力層 $\mathbf{x}(t)$ であり，第2層（競合層）の各ノードは入力層の次元に合わせてn個の要素を持ち，出力を視覚的にみるため2次元に配列されている． SOMの形成過程は競合層におけるノードの競合学習と自己組織化の2段階からなり，この2つの要素を取り入れることで競合層のノードは自己組織化され，2次元のSOMを生成する．

競合学習

競合層のノードiが時刻tで結合荷重 $w_i(t) = (w_{i1}, w_{i2}, \ldots\ldots, w_{in})$ をもち，外部から入力信号 $\mathbf{x}(t)$ が入ってきたとき，ノードiはこの入力信号を学習して次の時刻t+1には入力信号により近い結合荷重 $w_i(t+1)$ へ近づくように次式

$$w_i(t+1) = w_i(t)+\alpha(t)[\mathbf{x}(t)-w_i(t)] \quad (0<\alpha(t)<1) \tag{4-1}$$

に従って更新する．入力 $\mathbf{x}$ と w_i との距離が最小となる結合荷重 w_i を持つノードは勝者ニューロンと呼ばれる．

自己組織化

勝者ニューロンiの近傍のノードのセット $N_i(t)$ を考え，これらのノードには勝者ノードと協調することを促すように結合荷重 w_k は次式

$$w_k(t+1) = w_k(t)+\alpha(t)Y(t)[\mathbf{x}(t)-w_i(t)] \quad (0<\alpha(t)<1) \tag{4-2}$$

（ここで，$Y(t) = 1$ for k inside of $N_i(t)$，$Y(t) = 0$ for other k）

に従って更新される．近傍の拡がりの大きさを表す近傍関数 $N_i(t)$ は時間発展とともに狭くなるように与える．

4．2 化学構造式とSOM

これまで医薬品となる化合物の特徴を表現する顔として用いられてきたのは，2次元の化学構造式（Fig.10a）と3次元の立体構造表示（Fig.10b）である．Fig.10cには3次元の分子特性を表示する例として分子表面の静電ポテンシャルを示すがやや複雑である．[32] 新しいタイプの2D表示として，Kohonenのニューラルネットワークにより，数千にもおよぶ分子表面の多くの静電ポテンシャル値から2次元平面への射影により得られた自己組織化地図（SOM）をFig.10dに示した．[32] Fig.10aの化学構造式とFig.10dのSOMを比べてみると，どちらも2次元表示ではあるが，情報の中身はFig.10dの方がはるかに濃く，構造だけでなく特性についても知ることができ，この例では2次元平面の中に分子表面の特徴が抽出されている．

多くの効き目のあるエンドセリン受容体拮抗薬に存在するメチレンジオキシフェニル基は生体内でチトクローム P450と望ましくない代謝の相互作用をもつ．Anzaliら[35] はそのような相互作用が少なくかつメチレンジオキシフェニル基と同等の機能をもつ代用物を探すために，自己組織化ニューラルネットワークを使用してい

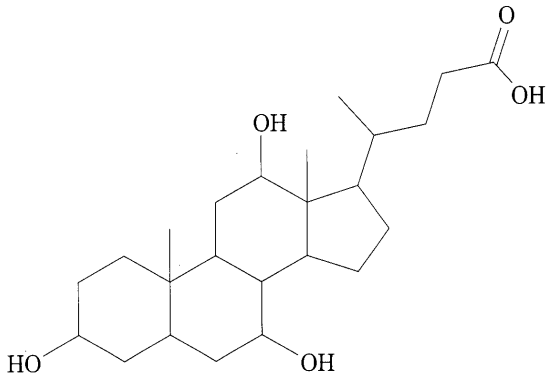


Fig.10a. 2D structure of Cholic Acid

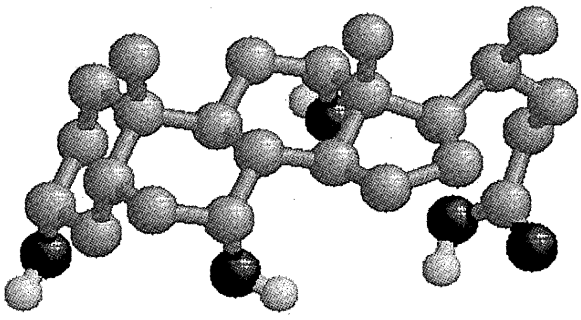


Fig.10b. 3D structure of Cholic Acid

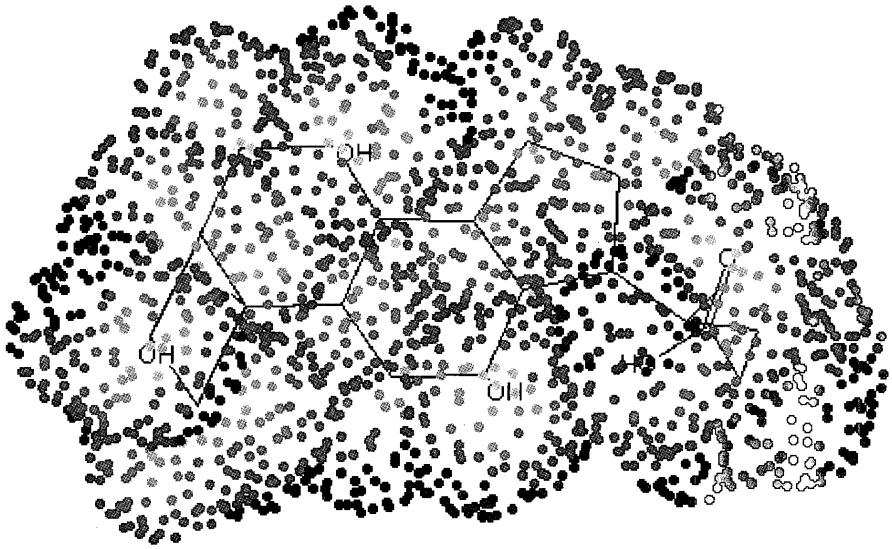


Fig.10c.　3D Van-der-Waals surface of electrostaic potential based on partial charges of the atoms

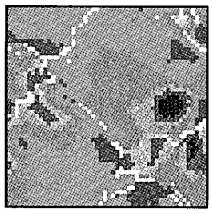


Fig.10d.　new 2D Self-Organizing Map for Van-der-Waals surface of electrostatic potential shown in Fig.10c.

くつかのエンドセリン受容体リガンドの静電ポテンシャルの特徴を分析した．フラグメントと官能基およびそれらに対応する自己組織化地図（SOM）の小規模なライブラリを作成した所，その中にメチレンジオキシフェニル基とよく似ているSOMが5通り存在し，特に類似度の高いベンゾチアゾール基を，メチレンジオキシフェニル基の代用物として選択することに成功している．

4．3　クラス分類問題

ノルボルナンおよびノルボルネンはChart 3に示すような基本骨格を有する．ノルボルナン誘導体の2位の置換基の立体配位とケミカルシフトとの関係を学習させ，未知誘導体の立体配位を予測させるという問題は，古くから線形学習機械法やクラスター分析法で解析され，ニューラルネットワークが導入されると同時に，BPの適用により予測精度が改良された問題である．[4] Fig.11がSOM計算[35]により得られた2次元マップであり，ノルボルナン誘導体38個に対応する勝者ニューロンと残りの362個のノード（図中の黒丸）が分布しているが，そのノード間の距離をグレーレベルで表現したものなのでグレーマップと呼ばれ，距離が大きいところは灰色が濃くなっている．*exo*型と*endo*型の誘導体はそれぞれがほぼグループ化されて分布しており，X1～X13（*exo*型）とD14～D25（*endo*型）の分布から，残りの誘導体（26～38）についてはマッピング位置からどちらのグループに属するかをほぼ正しく予測することができる．26がどちらかのグループに属するかの判断はやや微妙である．

また，32種類のカルボニル化合物をα位とβ位における4通りの解離反応に対応して，各誘導体のパラメータの違いから正しく分類予測する問題についてもSOMの適用により良好な結果が得られている．[36] これらの結果はSOMによるビジュアル化に基づくデータ解析により，BPによる解析で得られたものとほぼ同等の分類が可能であることを示している．

SOMにより良好な結果が得られる問題は，本

来，クラス間の分離がはっきりしている場合のはずで，上記の成功例は統計学的に言えばベイズ誤り確率（Bayes error）が極めて小さいデータ集合に対応すると思われる．[37] クラス間に重なりが存在するデータの場合はSOMによる分類では不十分になり，SOMを拡張して第3層に出力層を追加したネットワークである修正カウンタープロパゲーションネットワーク（MCP）が適している．[34] ただし，MCPでは，BPの場合と異なり，1層－2層間のノードの結合荷重への教師信号の影響はない．MCPは反応における試薬の役割分類と予測の問題に適用され成功している．[38] MCPの他に，自己組織化ネットワークにベイズ推定を取り入れた拡張版が役立つ可能性もある．

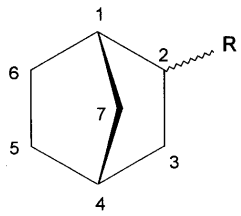


Chart 3. Structures of Norbornanes

マイトマイシンやアリルアクリロイルピペラジン類の等級活性分類については筆者らのSOMを使用したグレーマップによる分類では不十分な結果しか得られていない．これらに対しては，MCPの適用によっても大きな改善はみられず，BPによる計算結果も同じ程度に不十分であった[39] ことを考えると，‘次元の呪い’という言葉で指摘される，十分な解析をするにはサンプルデータが足りないデータの可能性が高い．[37]

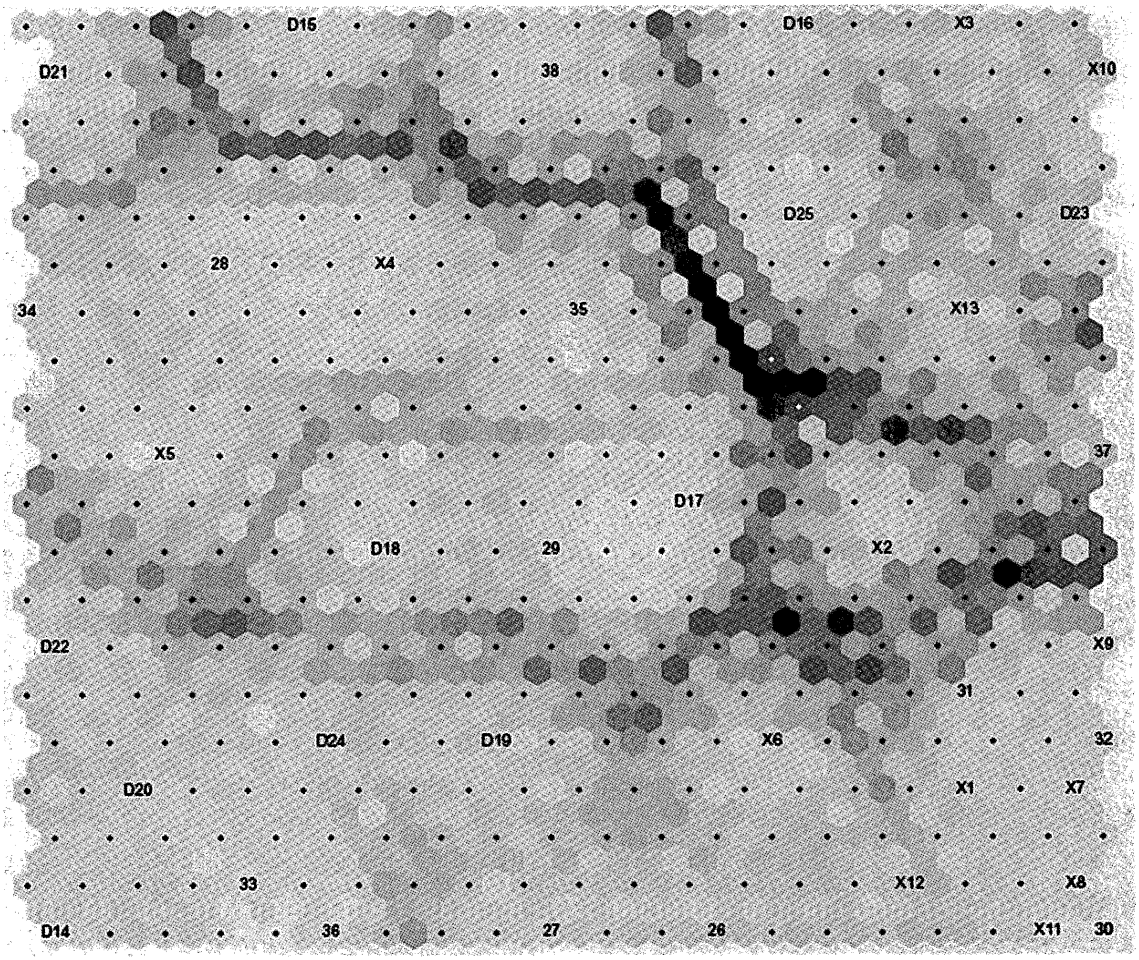


Fig.11. Kohonen's SOM calculated for 38 Norbornann/ Norbornene derivatives.
The label X represents “exo” type configuration and the label D does “endo” type.

4．4　欠損値SOMによるQSAR解析

BPやBRNNによるQSAR解析では，ある化合物の持っているn個の構造記述変数を入力層のノードに割り当て，出力層は活性値に対応する1つのノードに割り当てるのが普通である．SONNは教師データなしのアルゴリズムに従うため，この出力層ノードに直接対応するものは無い．そこで，SONNではサンプルデータ値に少しの欠損があってもそれを補間する能力にすぐれていることに着目し，n個の構造記述変数に活性値を加えた（n+1）次元データを入力層のノードに割り当てることで，SOMを使ったQSAR解析が可能となることを筆者らは見出した．[40] この場合，leave-one-out法を採用し，まず与えられたデータの中から1組のデータを除いてSOMを作成する．次いで，除いたデータの（n+1）番目の変数値（活性値）は欠損したものとして，このデータが既に作成したSOM内のどのノードに割り当てられるかを調べると(Fig.12)，そのノードの（n+1）番目の変数値から活性値を予測できる．

カルボキノンのQSAR解析を行なった結果，平均4.2%の誤差で活性値を予測することができており，生物活性値と予測値の相関係数は0.87と高いことがわかった．同様に，ベンゾジアゼピンの57誘導体のサンプルデータにつきQSAR解析を行なってみた所，物理化学パラメータ42変数を使用した計算では相関係数が0.5程度しか得られずSOMの手法によるQSAR解析は困難に思われたが，BP計算のpruningで見出された最適な物理化学パラメータ（6変数）のみを使用した計算を行なうと良好な相関係数～0.85が得られた．[41] これらの結果は，カルボキノンとベンゾジアゼピンでは欠損値SOMを使用したQSAR解析が，BPやBRNNによる場合に迫る予測精度をもつことを示している．ベンゾジアゼピンの場合に変数削減が必須であった事は，‘ρ =（パターン数/結合荷重の数）が数倍程度でニューラルネットワークがうまく働く’(rule of thumb）に合致している点で興味深い．このような構造記述変数が多いケースでは，重要度の低い変数の影響を排除することではじめてノード間の結合の強さを汎化能力の高い形で調整できたものと解釈できるだろう．QSAR解析には本来BPやBRNNが適しているが，SONNは教師なし学習アルゴリズムに従い多変数で非線形な構造をもつデータの特徴抽出やビジュアル化にすぐれた能力を示し，また，BPな

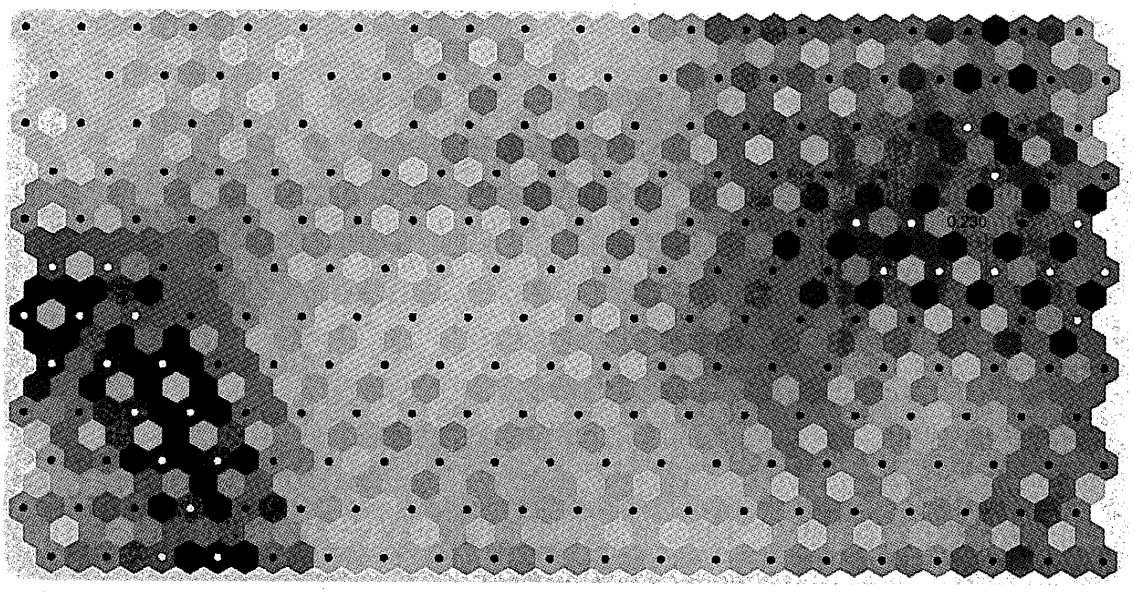


Fig.12. Kohonen's Self-Organizing calculation is made for No.6 Carboquinone with defective activity, and the winner neuron is settled. Predicted biological activity for No.6 is regarded as the value of 7-th dimentional component of it.

どに比べて計算の安定性が良いことが知られているので，BP や BRNN による QSAR 解析への補足的な手段として役立つことが期待できそうなので，今後は他の化合物についても調べたい．その際，一般には変数選択法を併用する必要があり，BRNN の ARD による適応度情報を用いるのが有用であると考えられる．

4．5 臨床医学への応用－甲状腺症患者分類

臨床医学における疾病の診断は分類問題であり，診断には外見的症状や問診のほか，多くの客観的データ，たとえば血清の構成成分，尿に含まれる代謝成分の生化学的検査値が用いられる．一般に，疾病は秩序ある体内の代謝系を乱し，生化学成分値が総合的に変化するため，診断には総合的な判断が必要とされる．しかし，人間では総合的解析は困難であり，さらに測定機器などの測定誤差や患者の個体差のため，現実の診断では特徴的な検査項目のみが注目され，他のデータは考慮されない．ニューラルネットワークは検査値と疾病の関係を総合的に学習し，測定値の変動や個体差を学習する機能を有し，さらに各疾病の経時変化を学習させることも可能である．

甲状腺症患者分類に自己組織化ニューラルネットワークによるクラスタリングを行ない作成したグレーマップを Fig.13 に，サモンマップを Fig.14 に，それぞれ示す．[42] 甲状腺の検査データの類

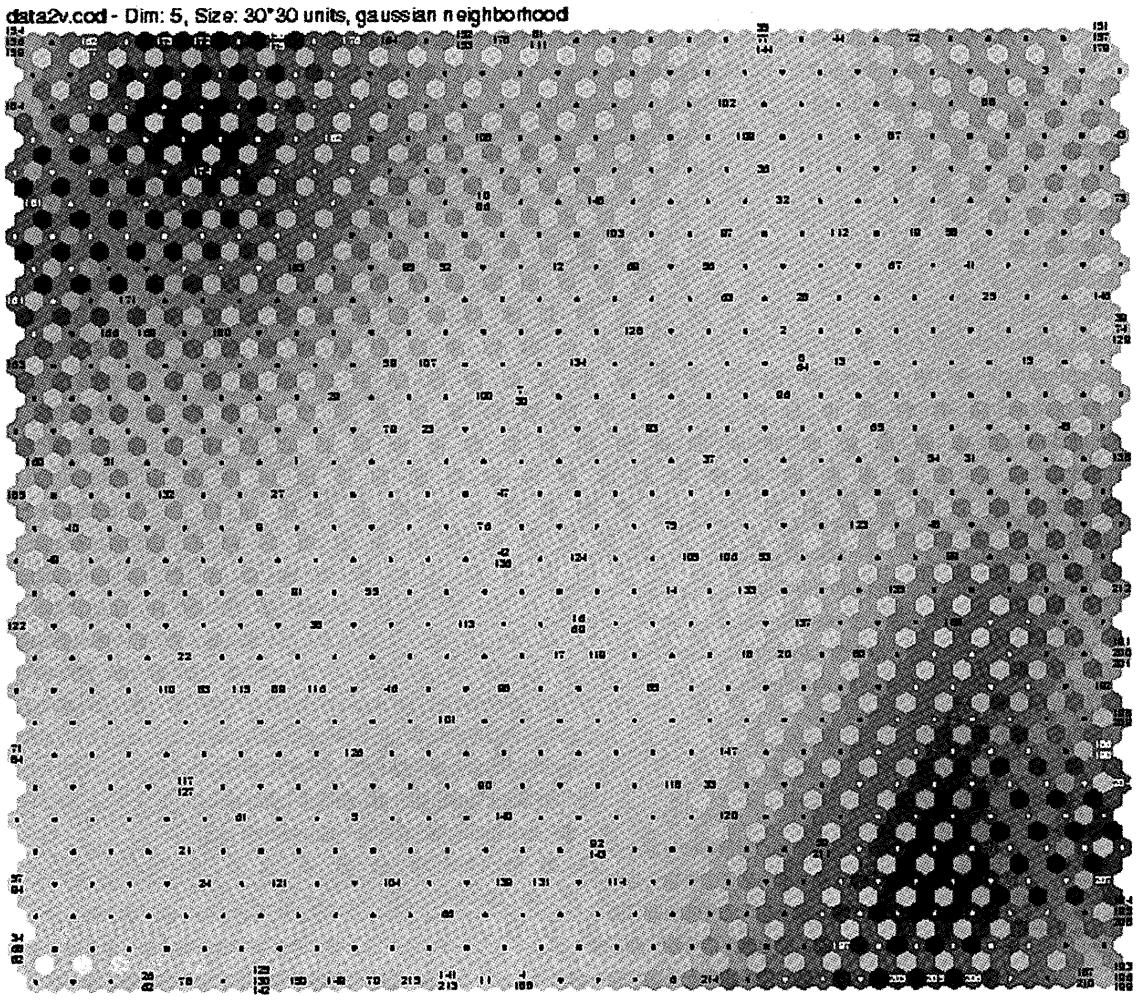


Fig.13. SOM for thyroid data set consisting of 215 patients form the same hospital. These individuals were divided into 3 groups that labeling 1 ～ 150 as euthyroid patients, 151 ～ 185 as patients suffering from hyperthyroidism, 186 ～ 215 as from hypothyroidism.

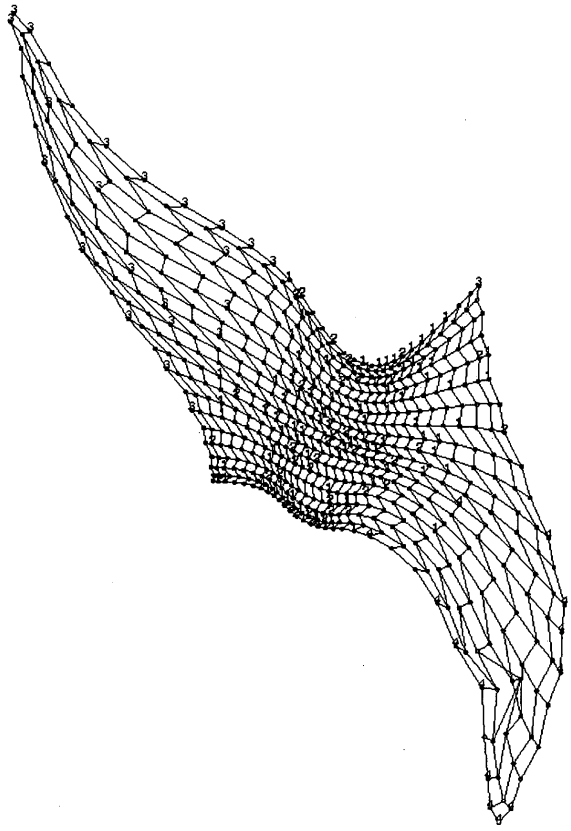


Fig.14. Sammon Map for thyroid patients data Label "1" and "2" corresponds to euthyroid, "3" to hyperthyroidism and "4" to hypothyroidism.

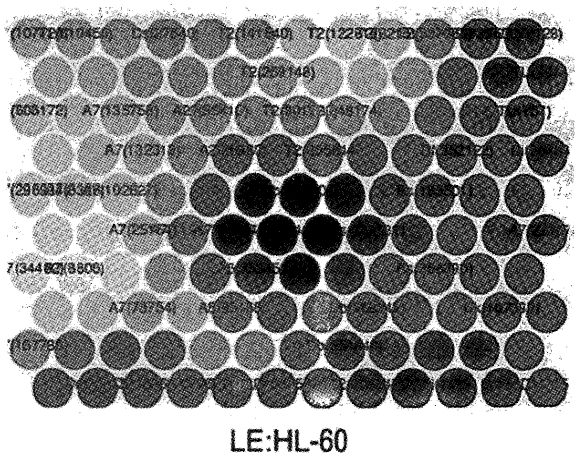


Fig.15. SOM visualizes activities of drugs for Leukemic cells. The lighter of the node, the more active of drug. Characters on node indicate pharmaceutical site of action.

似度が高い場合はサモンマップ上での距離が近くなり，類似度を明瞭によみとることができる．これは，TSH, T3RIA, T4, RT3U, DTSH の5項目からなるデータ[43]で，ラベル1～150は正常者，151～185は亢進症，186～215は低下症に相当する患者番号である．SOM では，亢進症，正常，低下症の患者グループに対応して，ほぼ3群にきれいに棲み分けたマップが得られた．亢進症と低下症の区別は完全であり，両者と正常の区別もほぼ可能である．しかし，境界ゾーンに位置する患者については分類が正しくないケースが数例みられた．これらの患者に対しては精密検査を実施するなどの必要性があると考えてもいいかもしれない．マップにより個々の患者の全体の中での位置づけが把握できることは，現在は正常でも甲状腺異常の予備軍にどの程度近いかを知るとか，診断における大きな誤りを回避する，等にも有効であると期待できる．

4．6　薬剤とがん細胞の相関解析へのSOMの応用

がん細胞と正常細胞における遺伝子発現変化と薬物感受性を対応づけて解析するために，例えば，米国国立がん研究所では60種類のがん細胞株（NCI60）について，きわめて多くの化合物に対する感受性および耐性と，多くの遺伝子の発現プロファイルをリンクさせた大規模データベースを構築している．通常，ほとんどが階層クラスタリング手法により解析されていたが，北野らは[44] Kohonen のニューラルネットワークにより NCI60 各細胞の種々の薬剤に対する感受性の違いによるクラスタリングを行った．Fig.15 に白血病細胞株（HL-60）に対する多数の薬剤をクラスタリングする SOM を示す．ノードが明るいほどその薬剤の HL-60 に対する反応性が高いことを意味する．SOM により，さまざまな側面からのより進んだ分類が可能になるものと期待されている．

5．おわりに

薬物の構造活性相関および医療・薬学分野での分類問題へのニューラルネットワークを応用した予測精度改善やデータのビジュアル化による特徴把握について述べた．予測精度改善とい

う点では，最近，サポートベクトルマシン（SVM）[45]というパターン認識手法が，タンパク質の分類問題への適用にかなりの成功をおさめた[46]ことで注目を集めており，医療・薬学分野での分類問題でもベイジアンニューラルネットワークを凌ぐ優れた手法になりうる可能性も期待される．サポートベクトルマシンは分類境界のデータだけを余裕をもって分離するというマージン最大化の方針による線形分類モデルであるが，カーネルトリックを用いて非線形のデータも取り扱えるように拡張されている．しかし，決定論的モデルであることから，確率を持ち込んだベイジアンニューラルネットワークほど優れた汎化性能を持つことは難しいかもしれない．[47]いずれにしても，BP の pruning，BRNN の ARD，SVM のマージン最大化，はどれも“ネットワークの自由度を低く抑えて汎化能力と解の安定性を高める”もので，本質的に重要である．自己組織化ネットワークの場合も優れた汎化性能を発揮するためには pruning が大切であるが，このネットワークは自身にその機能を持たないので，BP や BRNN での pruning の結果を反映させた SOM を作成すればよい．これらの手法はもっと多くの問題に応用し役立てることができるだろうし，それにより医療・薬学分野における様々なデータへの理解も進むだろう．ニューラルネットワークは基本的にパターン認識の手法であるから，今後は，医薬品情報の特徴抽出や視覚化，臨床医学における疾病診断の補助，などへの応用を期待したい．

REFERENCES

1) Fujita T. et.al., “Structure-Activity Relationships – Quantitative Approaches” (*Kagaku no Ryouiki zoukan* 122), Nankodo Co., 1979.
2) Fujita T. et.al., “Structure-Activity Relationships – Quantitative Approaches Ⅱ” (*Kagaku no Ryouiki zoukan* 136), Nankodo Co., 1983.
3) Sasaki S., Abe E., Takahashi Y., Takayama T., Miyasita Y., “*Kagakusya no tameno pataan ninsiki jyosetsu*”, Tokyo Kagaku Doujin Co., 1984.
4) Ichikawa H., “*Kaisougata nyurarunettowa-ku*”, Kyouritsu Co., 1993.
5) Zhang G, Berardi VL., *Health Care Manag Sci.* **1**, 29-37 (1998).
6) Asou H., Tsuda K., Murata N., “*pataan ninsiki to gakusyu no toukeigaku*”, Iwanami Co., 2003.
7) Usui S., Iwata A., Kyuma K., Asakawa K., “*Introduction and Practice of Neural Network*”, Corona Co., 1995.
8) Sato K., Nakagawa J., *Chem. Pharm. Bull.*, **45**, 107-115 (1997).
9) Rumelhart D.E. et. al., “*Paralell Distributed Processing*”, Vol.1, MIT Press, 1986.
10) Nakao H., Arakawa M., Nakamura T., Fukushima M., *Chem. Pharm. Bull.*, **20**, 1968-1974 (1972).
11) Yoshimoto M., Miyazawa H., Nakao H., Shinkai K., Arakawa M., *J. Med. Chem.*, **22**, 491-496 (1979).
12) Aoyama T, Suzuki Y, Ichikawa H., *J. Med. Chem*, **33**, 2583-2590 (1990).
13) Bishop C.M., “*Neural Networks for Pattern Recognition*”, Oxford University Press, 1995.
14) Maddalena DJ, Johnston GA., *J. Med. Chem.*, **38**, 715-724 (1995).
15) So SS, Karplus M., *J. Med. Chem.*, **39**, 5246-5256 (1996).
16) Winkler DA, Burden FR, Watkins JR., *Quant. Struct. -Act. Relat.* **17**, 14-19 (1998).
17) Turner JV, Cutler DJ, Spence I, Maddalena DJ., *J. Comput. Chem.*, **24**, 891-897 (2003).
18) Aoyama T, Ichikawa H., *Chem. Pharm. Bull.*, **39**, 1222-1228 (1991).
19) Cun, Y. L., Denker, J. S. and Solla, S. A., *Advances in Neural Information Processing Systems* **2**, 598-605 (1990).
20) Ishikawa M., *Neural Networks*, **13**, 1171-1183 (2000).
21) So SS, Karplus M., *J. Med. Chem.*, **39**, 1521-1530 (1996).
22) Randic'M., *J. Amer. Chem. Soc.*, **97**, 6609-6615 (1975).
23) MacKay, D. J. C., *Network: Computation in Neural Systems*, **6**, 469 (1995).

24) Neal R., M., "*Bayesian Learning for Neural Networks*", Springer, 1996.

25) Sato K., Nakagawa J., Matuzaki H., *Journal of Tohoku Pharmaceutical University*, **44**, 187-193 (1997).

26) Ajay, Walters W. P., Murcko M. A., *J. Med. Chem.*, **41**, 3314-3324 (1998).

27) Burden FR, Ford MG, Whitley DC, Winkler DA., *J. Chem. Inf. Comput. Sci.*, **40**, 1423-1430 (2000).

28) Burden FR, Winkler DA., *J. Med. Chem.*, **42**, 3183-3187 (1999).

29) Burden FR, Winkler DA., *Chem. Res. Toxicol.*, **13**, 436-440 (2000).

30) Lisboa PJ, Wong H, Harris P, Swindell R., *Artif. Intell. Med.*, **28**, 1-25 (2003).

31) Kohonen T., "*Self-Organizing Maps*", Springer, 2000.

32) Hemmer M. C., Gasteiger J., http://www.terena.nl/conferences/archive/tnc2000/proceedings/10B/10b5.html

33) Tetko I. V., Kovalishyn V. V., Livingstone D. J., *J. Med. Chem.*, **44**, 2411-2420 (2001).

34) Tokudaka H., Kisida S., Fujimura K., "*Jikososikikamappu no ouyou*", Kaibundou Co., 1999.

35) Anzali S., Mederski W. W. K. R., Osswald M., Dorsch D., *Bioorganic & Medicinal Chemistry Letters*, **8**, 11-16 (1998).

36) Sato K., Nakagawa J., Matuzaki H., *Journal of Tohoku Pharmaceutical University*, **48**, 125-131 (2001).

37) Ishii K., Ueda S., Maeda E., Murase H., "*Pataan ninsiki*", Ohmsha Co., 1998.

38) Satoh, H., Funatsu, K., Takano, K, Nakata, T., *Bull. Chem. Soc. Jpn.*, **73**, 1955-1965 (2000).

39) Aoyama T, Suzuki Y, Ichikawa H., *J. Med. Chem.*, **33**, 905-908 (1990).

40) Sato K., Hoshi K., Kawakami J., *Journal of Tohoku Pharmaceutical University*, **49**, 137-143 (2002).

41) Kawakami J., Hoshi K., Ishiyama A., Miyagishima S., Sato K., *Chem. Pharm. Bull.*, **52**, 751-775 (2004).

42) Hoshi K., Miyagishima K., Ishiyama A., Kawakami J., Nakamura H., Sato K., in preparation.

43) D. Coomans, M. Jonckheer and D. L. Massart, *Analytica Chimica Acta*, **103**, 409-415, (1979).

44) Muraki N., Kitano H., "Systematic Analysis of Cancer using very large database" in "*sisutemubaioroji no tenkai*" ed. by Kitano H., Springer-Verlag (Tokyo) 2001.

45) Cristianini N., Shawe-Taylor J., "*An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*", Cambridge University Press, 2000.

46) Zien A, Ratsch G, Mika S, Scholkopf B, Lengauer T, Muller KR., *Bioinformatics*, **16**, 799-807 (2000).

47) Liang F., *Neural Computation*, **15**, 1959-1989 (2003).