演　　題	神経回路網シミュレーションと構造活性相関 : 測定データが全て揃わない場合	
発　表　者 （ 所 属 ）	青山智夫、長嶋雲兵* (宮崎大学工学部、*産業技術総合研究所グリッドセンター)	
連　絡　先	宮崎大学工学部電気電子工学科 〒889-2192　宮崎市学園木花台西 1-1 TEL: 0985-58-7411, FAX: 0985-58-7411	
キーワード	Inverse problem, Neural networks, QSAR	
開発意図 適用分野 期待効果 特徴など	欠測のある不完全 QSAR データを解析できる多層階層型神経回路網シミュレータ A200 を開発した。A200 は 50%以上の欠測データを補間し QSAR 解析を可能にする。	
環　　境	適応機種名	IBM/AT 互換機
	O S 名	Windows 98, 2K
	ソース言語	Fortran 90 (Digital Fortran version 6)
	周辺機器	不要
流通形態 （右のいずれかに○をつけてください）	・日本コンピュータ化学会の無償利用ソフトとする ・独自に頒布する ・ソフトハウス、出版社等から市販 ・ソフトの頒布は行なわない ・その他　　○未定	具　体　的　方　法 Program の問い合わせ先 t0b217u@cc.miyazaki-u.ac.jp

1 研究目的

多階層型神経回路網はベクトルーベクトル変換器である。与えられた二種類のデータを「学習」と呼ばれる過程を経てその機能を発現する。最も良く使用される学習方式はback-propagation (BP) learningである。それは因果関係の不明な二種類の観測を、非線形であっても「必ず結びつける」が、それだけでは有用な機能ではない。危険でもある。同回路網の微分形 reconstruction learningが提示され、入出力間の因果関係を推定する指標が明らかになり、それによって同回路網は様々な用途、たとえば構造活性相関(QSAR)に適用されるようになった。しかし神経回路網は「データが無い」ということを学習できない欠点がある。Programmingを工夫し、無データ時そのデータ情報が伝搬する全神経間結合を固定し「学習不能」状態にすることは可能である。それはBP-learningにおいて「入力値=0」としたことと同じ結果となる。従って、今まで無データを正確に神経回路網で適切に取り扱う方法が無かった。理化学の測定で欠測は頻繁に起るので、それは同回路網の応用範囲を限定する。本論文の目的はその欠測を取り扱う神経回路網を提示することである。

2 QSARデータの特徴

欠測を取り扱う統計的方法：EM(Expectation and Maximization)-algorithmが研究されている[1]。それはRubinによれば「不完全データの問題を完全データの枠組で逐次的に解く」方法である。即ち：

(1) 欠測**X**が存在するとして不完全データを完全化し、問題を解く方法を定式化する。問題が解かれたとき、有意な量θを求めるとする。これをパラメータと考える。

(2) **X**の初期値を使ってその定式化された問題を解きθ 0を求める。

(3) θ 0を使用して初期の**X**を改良し、同様にしてθ 1を求める。

(4) その操作をθ iが収束するまで繰り返す。

ここで、QSARデータのとき(1-4)が妥当であるか検討する。(1-2)は問題ないが、(3)は有意であるか疑問である。θは一種の結果であって、化学反応系で「θを使って**X**を改良する」手段が存在するのか疑問である。それを統計処理上のalgorithmであると考え「その問題を問わない」ことにするとどうなるか検討する。神経回路網のQSAR処理を想定すると、(1)の定式化＝神経回路網の学習である。θは生理活性と神経回路網出力との偏差であろう。すると(3)は無意味になる。神経回路網の学習は偏差を極小化するための神経間の結合変更操作である。その操作の後にさらに入力を変更し偏差を極小化するのは危険である。統計学のように観測が確率分布$f(x|\theta)$から得られた数列と考えているのなら、観測に本来含まれていたはずの**X**とθを交互に(1-4)の最尤推定できるが、QSARデータはそういう観点から処理できないように思われる。

[1] 渡辺美智子、山口和範「EMアルゴリズムと不完全データの諸問題」ISBN4-8115-5701-8.

3 多層神経回路網補間

多階層型神経回路網(複数)で不完全データを解く方式は次のようになる。

(1) 生理活性データをソートする。その順序に従って構造データを並べ替える。生理活性データが欠測の場合はその順序位置を仮定する(仮定A)。完全なデータ部分だけを一旦ソートして、構造データの遷移状況を見て、それに合うような位置に欠測生理活性データを置く。ここでは神経回路網のpattern matching機能が使える。

(2) ソートされた一種類の構造データを教師データ、等差数列を入力データとして多階層型神経回路網(A)を学習する。ここは過学習に注意する必要がある。

(3) 神経回路網(A)から構造データの欠測部を補間する。

(4) 構造データが複数種類のとき、その種類数だけ(2-3)の操作を繰り返す。生理活性データに欠測があれば同じく(2-3)を繰り返す。その欠測位置の「仮定A」によって補間値が変化することに注意する。

(5) 以上測定データの完全化が成されたので、それを使用して第二の神経回路網(B)でQSARを実施する。

以上、EM-algorithmと違いiterativeな処理ではない。また仮定Aをself-consistenceな処理で決定することは困難である。

4 Carboquiones QSARへの適用

Carboquionesは抗癌剤で、公表されている諸データ[2]を見るとdummy dataのような特殊な説明変数も無い。そのデータに対し部分的に欠測したと仮定し**3**節の方法でQSAR計算を行った。仮定Aの問題が無いので理想的な結果であるが欠測率50%でもQSARに破綻は起きなかった。

[2] 構造活性相関懇話会「化学の領域122号薬物の構造活性相関」南江堂(東京), 1979.