応用力学論文集Vol.2,PP.83-90（1999年8月）　　土木学会

ニューラルネットワークを用いたデータマイニングによる非構造システムの同定

An Identification of Non-Structural System from Data Mining using Neural Network

須藤 敦史*・高須 光朗**・星谷 勝***
Atsushi SUTOH, Mitsuo TAKASU and Masaru HOSHIYA

* 正会員　博士(工学)（株）地崎工業土木技術部 主任研究員（〒105-8488 東京都港区西新橋2-23-1）
** 修士(工学) 武蔵工業大学工学部工学研究科　（〒158-8857 東京都世田谷区玉堤1-28-1）
*** 正会員　Ph. D.　武蔵工業大学工学部土木工学科教授　（〒158-8857 東京都世田谷区玉堤1-28-1）

In database, there has been a growing interest in efficient discovery, which is beyond the power of current data-processing functions, of interesting knowledge rule from huge database. The technology is called "data mining".

In this paper, we introduce data mining with a view to discuss applications of artificial life theory for data mining. The mechanism and major physical parameters for the generation of red tide are investigated within the framework of statistical data mining. Data mining means to discover objectively knowledge hidden in vast amount of data, and by means of neural network, data of Tokyo bay are analyzed. It is found that the usefulness of this data mining procedure for non-structural system identification.

Key Words : *data mining , neural network , non-structural system, identification*

1．はじめに

コンピュータ技術の発展により膨大なデータの蓄積や利用が可能となり，これらを効率良く活用するための解析手法やアルゴリズムの必要性が高まっている．

しかし，データには定性・定量データが混在するなど様式や形態が様々であったり，またそれらの相互関係が複雑であるため，現状のデータ処理技術ではデータの有効利用がなされていないが現状である．

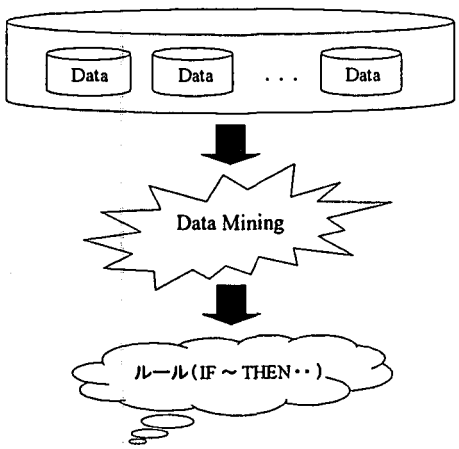
一方，従来のデータ処理・分析において確率・統計，機械学習，人工知能（Artificial Intelligence）やデータベース技術などが個別に用いられているが，より効率的でかつ高度なデータ処理・分析を行うためには，これらの技術等を融合してシステム化されたデータ処理および分析・解析手法が必要となる．

このような背景により，データベースに蓄積された膨大な生データから価値ある情報を発見することを目的とした「KDD（データベースからの知識発見）」あるいは「データマイニング（Data Mining）」に関する研究やその処理システムの構築が行われてきている[1),2)]．データマイニングとは膨大な生データの中に存在する隠れた知識や規則（ルール）を客観的に発見することであり，特にマーケティング分野において開発・応用事例が多く報告されている[3),4)]．

一方，このように多量の生データから隠れた知識や規則を発見するデータマイニングは，対象とする現象やデータ間の相関関係を観測値をもとに理解する逆解析的なアプローチと言えるが，データマイニングでは現象の関係式・支配方程式が明確になっておらず，加えて入力値が特定できない場合が多いため，特殊な逆問題に相当すると考えられる．

また，データマイニングに関する研究は確率・統計，機械学習や人工知能（AI）に関連する広範囲な研究成果を基礎として，様々な領域のデータベースに対して有効なアルゴリズムを構築し，データベースを効率良く利用することを総合的に明らかにする研究領域となっている[5)]．

そこで本研究は，ニューラルネットワークを用



図－1　データマイニングの概念図

いたデータマイニングの土木工学における非構造問題への適用として，東京湾で観測された水質観測（環境）データと赤潮発生との関係の解析を試みたものである．

ニューラルネットワークは，階層型と相互結合型に大別されるが，今回の解析では水質観測データと赤潮発生の関連性の把握・理解を目標としているため，数多く適用されている階層型ニューラルネットワークを採用して，ネットワークに重みを負荷して，その大きさで個々のデータ間と赤潮との関連性の強さを求めている．

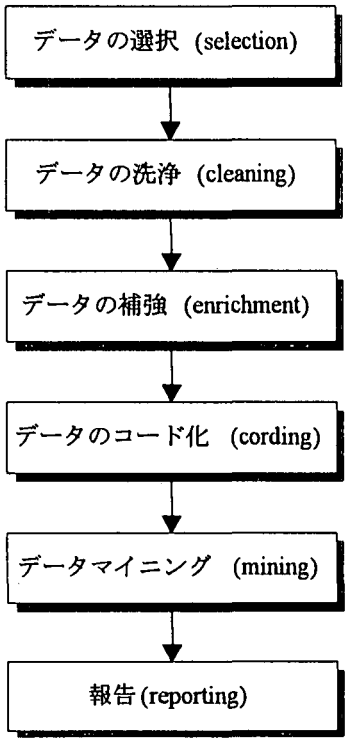
加えて，データマイニングの基本概念に従って，水質や環境に関する専門知識は全く参照せずに得られた生データのみから赤潮発生に関する要因や関連性の同定を試みている．

2．データマイニング

(1) データマイニングの概要[6)]

データマイニング(Data Mining)とは文字通り「データの発掘」であり，図-1の概念図に示すように膨大なデータの中の隠れた知識や規則を客観的に発見することである．このほか，データベースからの知識発見(Knowledge Discovery in Databases)，知識発掘(knowledge mining，knowledge extraction)，データ考古学(data archaeology)，データ浚渫(data dredging)などとも呼ばれている．

しかし，膨大なデータベースの中から何らかの情報を探索する研究は，従来より確率・統計，機械学習，人工知能やデータベース技術などで行われており，特に新しい研究分野ではない．しかし，近年データマイニングはこれらの技術を融合して



図－2　データマイニングのプロセス

システム化されたデータ解析技術の枠組みとして注目されており，この背景には以下に示すような要求が影響していると考えられる．

(a)データベースの発展

最近のコンピュータ技術の進歩により，膨大でかつ多様なデータの蓄積が進んでおり，加えてこれらデータベースの有効活用が求められている．

(b)理論・技術の統合

確率・統計，機械学習など，従来のデータ解析理論・技術は無関係あるいは重複して研究がなされているのが現状であるため，これらを統合・システム化してすべてを見通せるような新しい解析手法が求められている．

(c)技術のソフトウェア化

実際の大規模データ解析に適用するため，データマイニングにより統合・システム化され，かつ操作手順が簡単な汎用化，ソフトウエア化されたものが求められている．

（2）データマイニングのプロセス

一般的なデータマイニングにおける５段階のプロセスを図－２に示す．

(a)データの選択(selection)

事象と解析目標を設定し，基本となるデータベースから必要なデータを選定し，データマイニン

グ用のデータベースを構築する．

(b)データの洗浄(cleaning)

構築したデータベースから，ノイズや異常値と判別できるものを除去し，データをクリーンなものにする．ただし，この洗浄は前もって実行できるものもあれば，データのコード化や知識の発見の段階になってから実行する場合もある．

加えて，連続データの離散化処理などの作業もこの段階に含まれる．

(c)データの補強(enrichment)

新たなデータやマイニングにおいて有用なデータを追加する．この段階で他のデータベースとの結合を考慮した構造形式にする．

(d)データのコード化(cording)

データをマイニングしやすい形式に変換する．

(e) データマイニング(mining)

前処理が済んだデータベースから知識（ルール）の発見を行う．この段階がデータマイニングにおいて最も重要な段階である．

(f)報告(reporting)

データマイニングによって得られた知識や規則をグラフなどで整理し，分析結果としてまとめるものである．

ここで，一般的にデータマイニングはデータの解釈や方法によって「仮説検証型」と「仮説生成型」に分けられる．前者は仮説をデータによって検証することを主な目的とし，後者はデータを単純な表現形式に変換して隠れていた法則や規則を発見することを目的としている．しかし，両者とも従来の解析手法に比べて「状況の予測」より「結果の解釈」に重点を置いているところが特徴である．

（3）データマイニングに用いる解析理論

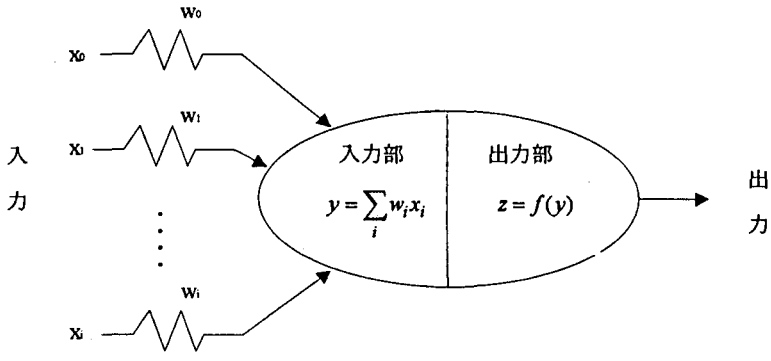
データマイニングは，大量なデータの中から隠れていた情報を見い出す解析理論・技術であるなら，基本的にどのような手法を用いても構わない．

したがって，データ解析の目的に応じて様々な理論や手法を単独あるいは複数組み合わせて用いることができる．つまりデータマイニングは‘データの解析技術・手法’というより，むしろ‘データ解析における概念・考え方’であると言える．

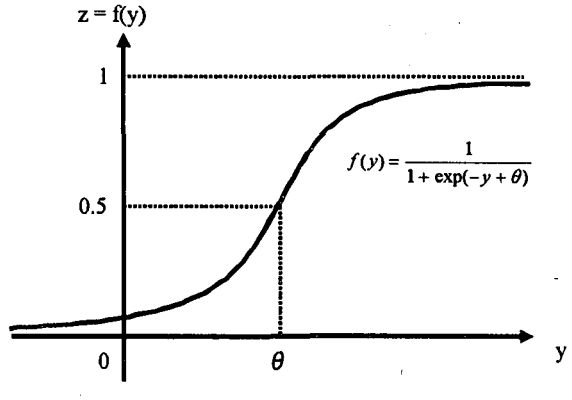
3．ニューラルネットワーク[7),8)]

ニューラルネットワーク(Neural Network)は人間の脳の構造を工学的にモデル化したものであり，主に逆問題，予測問題や組合せ最適化問題などに用いられている．一般にニューラルネットワークは，階層型と相互結合型に大別されるが，階層型ニューラルネットワークは基本的には教師データによりニューロン間の結合（重み）を変化させ，対象とする事象の学習を行うものである．

本研究では，この階層型ニューラルネットワークを用い，重みを変化させることより，各観測データと赤潮との関連性の把握を試みている．



図－３　ニューロンのモデル化



図－４　シグモイド関数

(1)ニューラルネットの情報伝達

ニューラルネットワークのニューロン（神経細胞）をモデル化したものを図－３に示す．

ここでニューロンは信号を受け取る‘入力部’と信号を発信する‘出力部’に分かれている．入力部では前のニューロンからの入力の総和を計算し，その結果を出力部においてでシグモイド関数などにより出力が判定され，次のニューロンへと情報が伝達される．

ここで図－３において x_i ：前細胞の出力値，w_i ：結合の強さ（重み）および $f(y)$ ：出力関数，θ ：しきい値である．

加えて，図－4には，本解析で用いた出力を判断するシグモイド関数を示す．

(2)ニューラルネットワークにおける学習

本研究では赤潮発生と個々の水質観測データ間の関連性の抽出を目的としているため，階層型ニューラルネットワークを採用し，個々のデータ間と赤潮との関係の強さはネットワーク結合（重みの大きさ）で定義する．

ここでニューラルネットワークにおける学習は，各ニューロンにおける重みを変化させ，入力・出力間に最適なネットワークを構築することをである．本解析では出力値（赤潮の発生）に対する入力値（各水質観測データ）として，各ニューロンにおける重みの大きさを求めている．

以下に本解析で用いた重みの学習法（バックプロパゲーション法：BP 法）の概略および BP 法を拡張した仮想インピーダンス法と成長抑制学習を記述する[9]．

BP 法は，階層型ニューラルネットワークにおいて代表的な学習方法であり，誤差逆伝播法とも呼ばれている．BP 法ではネットワークの出力値と教師信号との誤差が小さくなるように各ニューロンの結合重みを調節するものである．

いま，ある時刻 t の入力信号 p に対する出力の絶対誤差を式(1)と定義する．

$$E_p(t)=\frac{1}{2}\sum_i\left(T_{pi}-O_{pi}\right)^2 \qquad (1)$$

ここで T_{pi} は入力信号 p に対する i 番目出力層の教師信号，O_{pi} は T_{pi} に対応するネットワークからの出力値である．

この出力誤差を式(2)に示すように P 個の入力信号について平均し，ネットワーク全体の出力誤差を算出する．

$$E(t)=\frac{1}{P}\sum E_p(t) \qquad (2)$$

そこで，式(3)を用いて式(2)の平均出力誤差が最小となるように各ニューロンの重み（結合）を調節する．

$$\Delta W_{ij}^k(t)=-\varepsilon\frac{\partial E(t)}{\partial W_{ij}^k(t)}+\alpha\Delta W_{ij}^k(t-1) \qquad (3)$$

W_{ij}^k は k 層 i 番目，k-1 層 j 番目ニューロンの結合の重みである．

ここで，式(3)におけるパラメータ α は出力誤差の振動を減衰させる効果を有しており，解の安定に効果を示すが，学習速度の低下や解が局所解に滞留するおそれがある．

そこで，BP 法を拡張した仮想インピーダンス法を用いて，事象に対する高速学習を行う．

この方法は式(4)に示すように(3)式にパラメータ β を含む項を追加して各ニューロンの重みを調節する．

$$\Delta W_{ij}^k(t)=-\varepsilon\frac{\partial E(t)}{\partial W_{ij}^k(t)}+\alpha\Delta W_{ij}^k(t-1)+\beta\Delta W_{ij}^k(t-2) \qquad (4)$$

ここでパラメータ β はネットワークの学習中に重みに強制振動を与え，このことにより式(1)を局所解（ローカルミニマム）から脱出させる効果を有している．

次に，成長抑制学習は仮想インピーダンス法をさらに拡張するものである．

BP 法や仮想インピーダンス法では多くの重みが複雑に結合してしまうため，ニューロン間の重みの大きさを特定することが困難である．しかし，成長抑制学習では重みの成長に抑制をかけて強い（重みが大きい）結合だけが生き残るように調節するために，ニューロン間の結合が明確になる特徴を有している．

成長抑制学習は式(5)に示すように成長側に対して抑制項 S を追加したものであり，これにより各ニューロンの重みの大きさを調節する．

$$\Delta W_{ij}^k(t)=-\varepsilon\frac{\partial E(t)}{\partial W_{ij}^k(t)}+\alpha\Delta W_{ij}^k(t-1)+\beta\Delta W_{ij}^k(t-2)+S \qquad (5)$$

この式(5)における成長抑制項 S は，式(6)で定義される．

$$S=-s\frac{1}{m-1+1}\mathrm{sgn}\left(W_{ij}^k(t)\right)\left\{\sum_{l=1,\neq j}^{m}\left|W_{il}^k(t)\right|+\left|\theta_i^l\right|\right\} \qquad (6)$$

s：成長抑制係数，　m：k-1 層のユニット数

sgn (x)：x<0 のとき-1，x=0 → 0，

x>0 → +1 となる関数

したがって，パラメータ s は重みの成長抑制の効果があり，s を大きくするほど弱干精度は落ちるものの，各ニューロンの結合形態は簡潔化され，得られる規則は単純な表現形式となる．

表－1　水質（環境）観測項目

		説明
A	気温	
B	水温	
C	透明	
D	pH	
E	COD	化学的酸素要求量(mg/l)
F	DO	溶存酸素量(mg/l)
G	T-P	全ﾘﾝ(mg/l)
H	PO4-P	ﾘﾝ酸態ﾘﾝ(mg/l)
I	T-N	全窒素(mg/l)
J	NH4-N	ｱﾝﾓﾆｱ態窒素(mg/l)
K	NO2-N	亜硝酸態窒素(mg/l)
L	NO3-N	硝酸態窒素(mg/l)
M	SAL	塩分(mg/l)
N	Chl-a	ｸﾛﾛﾌｨﾙa(mg/l)
O	赤潮	赤潮発生の有無

表-3　属性の簡略化

		平均値未満	平均値以上
A	気温	A_1	A_2
B	水温	B_1	B_2
C	透明	C_1	C_2
D	pH	D_1	D_2
E	COD	E_1	E_2
F	DO	F_1	F_2
G	T-P	G_1	G_2
H	PO4-P	H_1	H_2
I	T-N	I_1	I_2
J	NH4-N	J_1	J_2
K	NO2-N	K_1	K_2
L	NO3-N	L_1	L_2
M	SAL	M_1	M_2
N	Chl-a	N_1	N_2

※ 赤潮あり:O_1　赤潮なし:O_2

表－2　基本的な統計量

		全データ			赤潮ありデータ			赤潮なしデータ		
		最小	平均	最大	最小	平均	最大	最小	平均	最大
A	気温	11.0	22.07	33.5	15.0	22.33	30.5	11.0	22.02	33.5
B	水温	11.5	20.88	30.2	15.0	21.10	28.5	11.5	20.84	30.2
C	透明	0.4	2.18	9.0	0.4	1.19	2.5	0.4	2.36	9.0
D	pH	7.2	8.16	9.0	7.7	8.39	9.0	7.2	8.11	8.7
E	COD	1.4	4.58	20.0	4.0	7.31	20.0	1.4	4.07	7.9
F	DO	4.2	8.36	16.1	7.2	10.82	16.1	4.2	7.90	13.4
G	T-P	0.0320	0.14481	0.7800	0.0460	0.19813	0.7800	0.0320	0.13490	0.5300
H	PO4-P	0.0010	0.07493	0.4500	0.0010	0.04719	0.3600	0.0010	0.08008	0.4500
I	T-N	0.380	1.8844	6.300	0.890	1.9913	4.500	0.380	1.8645	6.300
J	NH4-N	0.010	0.5260	3.000	0.010	0.2947	2.200	0.010	0.5690	3.000
K	NO2-N	0.0050	0.07482	0.3700	0.0100	0.06438	0.1700	0.0050	0.07676	0.3700
L	NO3-N	0.0020	0.65311	3.1000	0.0500	0.46957	1.5000	0.0020	0.68720	3.1000
M	SAL	1.4900	25.09940	32.2100	13.1900	25.89702	30.6800	1.4900	24.95123	32.2100
N	Chl-a	0.4	35.43	620.0	25.0	130.51	620.0	0.4	17.77	87.0

4．非構造システム（赤潮発生要因）の同定

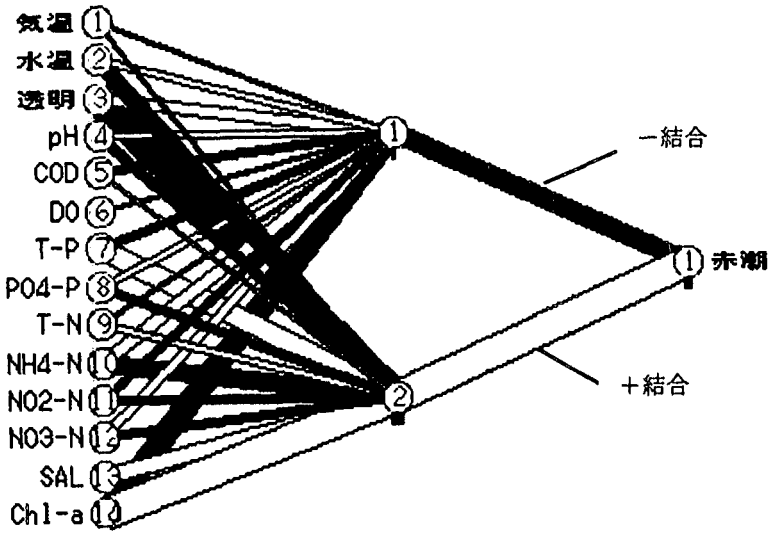
(1) 赤潮問題の設定

データマイニングに用いるデータベースは表－1に示す項目で観測されている東京湾東部における 10 観測点の水質データを用いている．ここで表－1においてA～Nは数値属性，赤潮に関する項目Oはブール属性（0-1）のデータである．

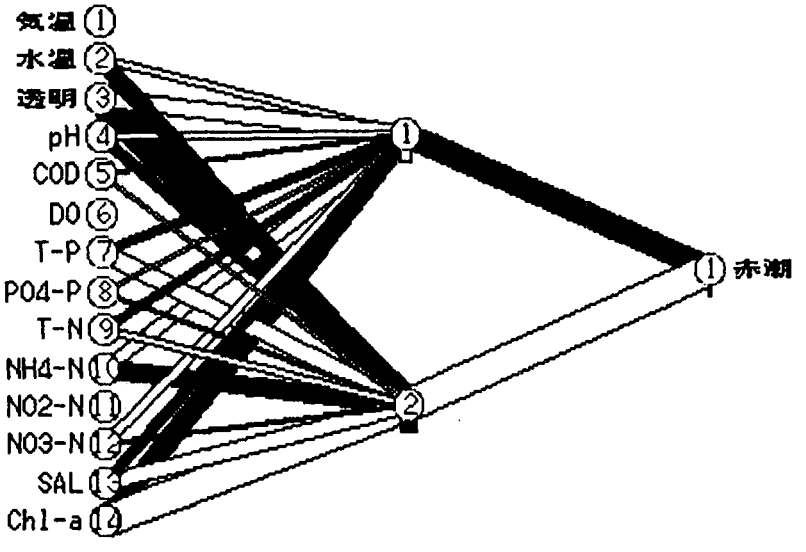
加えて，水質データは 10 観測点とも 1988～92 年の 5 年間における 4～9 月（比較的赤潮が発生しやすい時期）に観測されたものであり，データ総数は 300 である．

ここで観測データの基本的な統計量として，各観測データにおける最大・最小，平均値を表－2に示す．なお，赤潮発生頻度は全 300 データの中で 47 回であった．また，関連性解析の簡素化のために，各観測データの属性を表－3に示すように平均値以上・以下のブール属性（0-1）としている．

そしてニューラルネットワークにおいて個々の観測項目と赤潮とのネットワークの結合（重みの大きさ）度を求め，この結合度により各項目と赤



図－５　BP 法による学習結果



図－６　成長抑制学習の結果

潮との関連性を単純な「If ～ then …」形式のルールとして抽出することとする.

(2) ニューラルネットワークによる関連性の分析

ニューラルネットワークを用いたデータマイニングにより水質観測データと赤潮発生に関する関連性の分析を行う.

(a)解析モデル・学習データ

ここでニューラルネットワークは中間層１層（ノード数 2）を有するモデルを用いており，表－１における気温～Chl-a までの 14 入力値とし，赤潮発生の有無を出力値としている．また成長抑制係数は予備解析により求めている.

ここで入力値は簡素化と各観測数値の単位やディメンジョンに統一性がないため，各観測項目ごとに平均値未満（小）のデータを「0」，平均値以上（大）のデータを「1」としている.

(b)ネットワークによる関連性の学習

まず，バックプロパゲーション法により学習を行った結果を図－５に示す.

図－５より，各ニューロン間の結合が複雑になりすぎて観測項目と赤潮発生の関連性を判別することが難しい.

そこで，ネットワークの構築がある程度進んだ段階で，重みの学習方法を成長抑制学習に切り替える操作を行う.

ここでネットワーク構築中に学習方法を切り替えた結果を図－６に示す.

図－６より，成長抑制学習の効果により，各ニューロン間の結合は簡素化が計られているが，まだ単純なルールには至っていない.

ここで，図中のネットワークにおいて「＋結合」は各項目における平均値以上の属性と赤潮との関連性を意味しており，また「－結合」は平均値以下の属性との関連性の強さである.

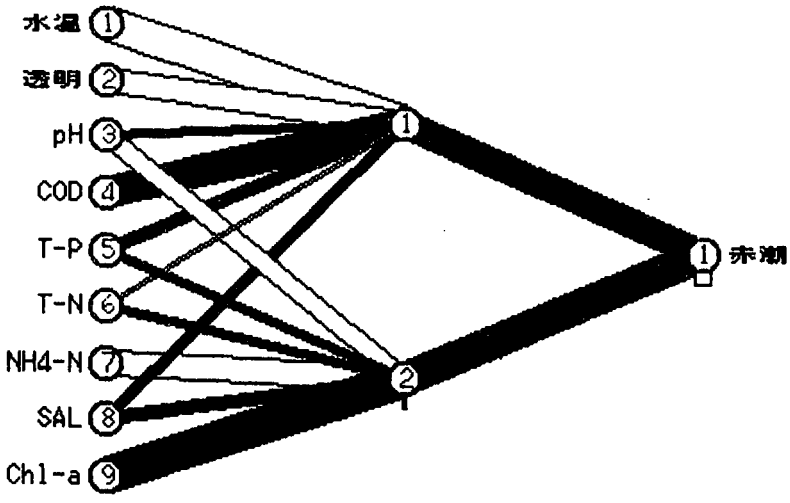
加えて，関連性の強さはニューロン間の結合の太さで表している.

(c)ネットワークの再構築

さらに各ニューロン（観測項目）間の結合の簡素化，明確化を計るために，ネットワークの構築が進んだ段階で，重みの小さくなった結合（平均結合以下）の削除を行う.

この操作により関連性の強いネットワークのみの再構築結果を図-7 に示す.

図－７より，ネットワーク再構築の結果，重みが小さくなった「気温」，「DO」，「PO4-P」，「NO2-N」「NO3-N」の項目が削除され，ネット



図－7　最構築後の学習結果

表-4　解析結果

属性		結合の強さ	結合状態	ルール
B	水温	強	(+)(−)	「0:小」→「1:赤潮あり」 (「1:大」→「0:赤潮なし」)
C	透明	中		
J	NH4-N	中		
E	COD	強	(−)(−)	「1:大」→「1:赤潮あり」 (「0:小」→「0:赤潮なし」)
G	T-P	弱		
I	T-N	弱		
M	SAL	弱		
N	Chl-a	強		
D	pH	弱	(+)(−)	「0:小」→「1:赤潮あり」 (「1:大」→「0:赤潮なし」)
		弱	(−)(−)	「1:大」→「1:赤潮あり」 (「0:小」→「0:赤潮なし」)

ワーク結合の簡素化がなされている.

加えて,「水温」,「透明」,「NH4-N」は平均値以下が,「COD」,「T-P」,「T-N」,「SAL」,「Chl-a」では平均値以上が,赤潮が発生との関連性が高く,加えて結合の強さから相関の度合いが評価できる.

したがって,ニューラルネットワークを用いたデータマイニングにより得られたルールをその関連性の大きい順に記述すると以下のようになる.

第１ルール：「水温：小」,「COD：大」,
「Chl-a：大」
第２ルール：「透明：小」,「NH4-N：小」
第３ルール：「T-P：大」,「T-N：大」,
「SAL：大」

ここでpHに関しては平均値以上・以下,両方の関連性が抽出されているが,これは他の項目との連動して抽出されたもと考えられるため,今回のルールからは除いている.

5．結　論

ニューラルネットワークを用いたデータマイニングの有効性の検討を行い,同時に東京湾において実際に得られた水質観測データと赤潮発生との関連性の解析を行った結果,以下に示すような結果が得られた.

(1) 各観測項目を平均値未満と平均値以上のブール属性データに分け,同時に重みの小さくなった結合を削除する成長抑制学習を適用することにより,ニューラルネットワークの簡素化が図れ,赤潮発生との関連性の強いと思われる観測項目の抽出が簡単になった.

(2) 赤潮の発生と関連性が抽出された観測項目とその強さは

第１ルール：「水温：小」,「COD：大」,
「Chl-a：大」
第２ルール：次に「透明：小」,「NH4-N：小」
第３ルール：「T-P：大」,「T-N：大」,
「SAL：大」

のような結果となった.

(3) 以上より,ニューラルネットワークを用いたデータマイニングの有効性が確認された.

また,今後の課題として以下が挙げられる.

(1) 実データを用いたデータマイニングでは効率的な前処理が重要であり,簡単に扱える汎用的な前処理手法の検討が必要となる.

(2) 関連性が抽出された項目の時系列な分析や検討が必要となる.

(3) 本来,環境問題は入力・出力と分けられないため,今後は相互結合型ニューラルネットワークなどによるを入・出力を規定しない解析が必要である.

参考文献

1) Pieter Adrians・Dolf Zantinge 著，山本英子・梅村恭司 訳：データマイニング，共立出版，1998.
2) 河野博之：データベースからの知識発見の現状と動向，人工知能学会誌，Vol.12，No.4，pp.496-504，1997.
3) 沼尾雅之，清水周一：流通業におけるデータマイニング，人工知能学会誌，Vol.12，No.4，pp.528-535，1997.
4) 中林三平：データマイニング 価値ある情報を掘り当てる，NIKKEI COMPUTER，pp142-147，1996.
5) 喜連川優：データマイニングにおける相関ルール抽出技法，人工知能学会誌，Vol.12，No.4，pp.513-520，1997.
6) 寺野隆雄：KDD ツールの動向と課題，人工知能学会誌，Vol.12，No.4，pp.528-535，1997.
7) 萩原将文：ニューロ•ファジイ•遺伝的アルゴリズム，産業図書，1995
8) 市川 紘：階層型ニューラルネットワーク，共立出版，1993.
9) NEUROSIM/L light，富士通，1996

(1999.4.23 受付)