(5) ニューラルネットワークを用いたデータマイニングの環境問題への適応について

An Adaptation of Environment Issue from Data Mining using Neural Network

須藤 敦史*・佐藤 大介・星谷 勝*****
Atushi SUTOH, Daisuke SATOH, Masaru HOSHIYA

* 博士(工学) ㈱地崎工業 技術開発部 主任研究員 (〒105-8488 東京都港区西新橋2-23-1)
** 武蔵工業大学大学院 工学研究科土木工学専攻 (〒158-8557 東京都世田谷区玉堤1-28-1)
*** Ph.D. 武蔵工業大学教授 工学部土木工学科 (〒158-8557 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 is analyzed. It is found that the usefulness of this data mining procedure for adaptation of environment issue.

Key Words : data mining, neural network, environment issue

1. はじめに [1) ~ 5)]

コンピュータ技術の発展により，様々な種類のデータを大量に蓄積することが可能となってきている．しかし，今のところ，定性・定量データが混在し，データの形式は様々であり，さらに，それらの相互関係が複雑である場合が多い．したがって，現状のデータ処理技術では所有のデータを有効に活用しているとはいえない．

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たが，より有効的なデータ処理・分析を行うにはそれらの技術を融合し，システム化されたデータ処理および分析・解析手法の確立が必要である．

このような背景により，「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)」が注目されている．データマイニングとは，膨大なデータの中に存在するはずの隠れた知識や規則（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で多くの成功事例が報告されている．また，製造業や医療の分野，あるいは土木（建設）分野への応用についても数々の研究がなされている．しかし，未だ発展途上の段階にあり，各分野での研究成果が期待されている．

また，データマイニングとは，観測をもとにして対象とする現象やデータ間の相関関係を発見・把握する逆解析的なアプローチといえる．しかし，データマイニングでは現象の関係式・支配方程式が明確ではなく，加えて入力値が特定できないため，特殊な逆問題に相当すると考えられる．

そこで本研究では，ニューラルネットワークを用いたデータマイニングの土木工学，特に環境問題への適用に関する研究を行う．実データへの適用性を検討するため，東京湾で実際に観測された環境（水質）データを用いて赤潮発生要因との因果関係の解析を試みる．また，データマイニングの基本概念に基づき，水質に関する専門知識は全く参照せず，得られたデータのみから赤潮発生のメカニズムを分析することが目的である．

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，本研究では水質観測データと赤潮発生の関連性の発見・把握を目標としているため，適用例の多い階層型ニューラルネットワークを採用し，ネットワークの重みの大きさから個々のデータ間と赤潮発生との関連性の強さを算出している．

2. データマイニング [6)]

2.1 データマイニングの概要

データマイニング(Data Mining)とは，文字通り解釈すれば「データ発掘」であり，膨大なデータの中に存在する隠れた知識・規則（ルール）を客観的に発見することである（図-1）．

また，データマイニングは，データベースからの知識発見(KDD: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．

確率統計，機械学習やデータベース技術などデータから有用な情報を抽出する研究は以前から行われており，特に新しい研究分野ではない．しかし，データマイニングはこれらの技術を融合してシステム化されたデータ解析技

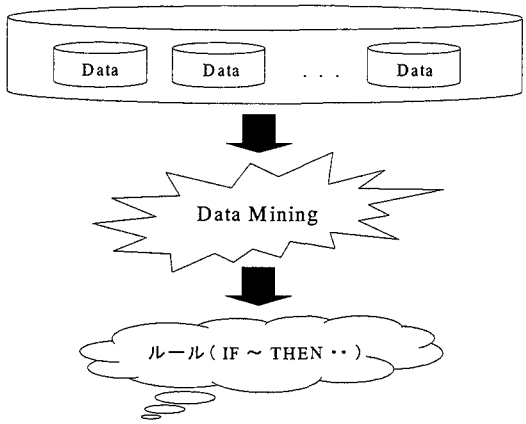

図-1　データマイニングの概念

術の枠組みとして注目されている．この背景には以下のような要求が影響していると考えられる．

(1)　膨大なデータの有効活用

近年のコンピュータ技術の発展により，膨大かつ多種多様のデータが蓄積されていく反面，それらデータを有効に活用することが求められている．

(2)　理論・技術の統合

確率統計，機械学習，データベース技術などのデータ解析理論・技術は無関係，あるいは重複して研究がなされているため，これらを融合し，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．

(3)　技術のソフトウェア化

成熟しつつある人工知能やデータベース技術をデータマイニングにより統合し，ソフトウェア化したうえで実際の大規模なデータ解析に適用することが求められている．

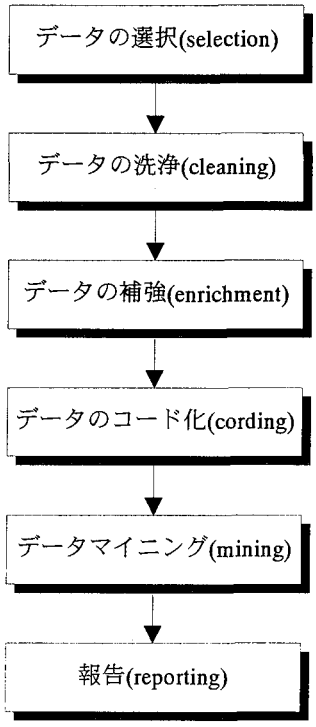

図-2　データマイニングのプロセス

2.2　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．

(1)　データの選択(selection)

解析目標を設定し，必要なデータを選択する．これにより，データマイニング用のデータベース（操作データベース）が構築される．

(2)　データの洗浄(cleaning)

構築されたデータベースからノイズや異常値を除去し，データをクリーンなものにする．この洗浄は，前もって実行できるものもあれば，コード化やマイニングの段階で誤りを発見してからでないと実行できない場合もある．また，連続データの離散化などの作業もこの段階に含まれる．

(3)　データの補強(enrichment)

新たに有用であると思われるデータを追加する．

(4)　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．

(5)　データマイニング(mining)

前段階までの処理が済んだデータベースから知識・規則（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といえる．

(6)　報告(reporting)

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

データマイニングは「仮説検証型」と「仮説生成型」に大別され，データマイニングの方法やデータの解釈方法が原理的にそれぞれ異なる．前者は利用者がもつ仮説を与えられたデータによって検証することが目的である．一方，後者はデータを操作することでより単純な表現形式に変換し，隠れていた法則を発見することが目的である．しかし，両者とも従来の手法に比べると「新しい状況を予測すること」よりも「結果をいかに解釈するか」に重みが置かれているのが特徴である．

2.3　データマイニングに用いる解析理論

データマイニングを行う際，多量なデータの中から隠れている情報を発見する解析理論・技術であるならば，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で，あるいは複数を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‘データ解析技術・手法’というよりも，むしろ‘データ解析における考え方’であるといえる．

3　ニューラルネットワーク [7) ~ 10)]

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．

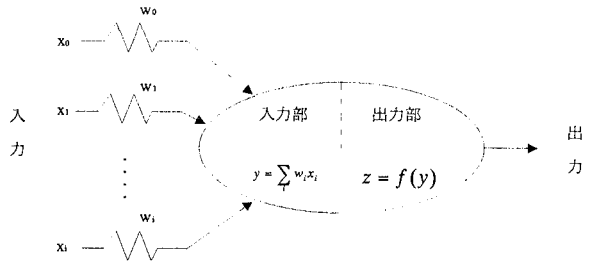

図-3　ニューロンのモデル化

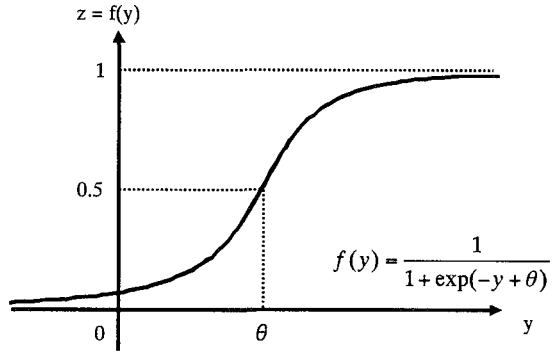

図-4　シグモイド関数

3.1　情報伝達の仕組み

図-3 は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である．ニューロンは，他のニューロンからの情報を受け取る‘入力部’と，他のニューロンへと情報を発信する‘出力部’に分かれている．入力部では前細胞からの入力の総和を計算する．その結果を出力部で判別し，出力が決定されて次の細胞へと情報が伝達される．図-4 に示すシグモイド関数は，出力関数の代表的なものである．図-3 において，x_i：前細胞の出力値，w_i：結合の強さ（重み），$f(y)$：出力関数，θ：しきい値である．

3.2　階層型ニューラルネットワークによる分析

ニューラルネットワークにおける学習とは，ニューロン間の結合の重みを変化させ，最適なネットワークを構築することである．本研究では個々の水質データ間と赤潮発生との関係の強さはネットワーク結合（重みの大きさ）で定義している．また，この学習には‘教師あり学習’と‘教師なし学習’があり，一般に階層型ニューラルネットワークは前者の教師あり学習を適用している．教師（信号）とは，最も望ましい出力値，すなわち入力値に対する実際の出力値のことである．以下に，代表的な学習方式であるバックプロパゲーション法（BP 法），さらに BP 法を拡張した仮想インピーダンス法，成長側抑制学習について示す．

バックプロパゲーション法（BP 法）は，階層型ニューラルネットワークの最も代表的な学習方式であり，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をフィードバックし，誤差が小さくなるように結合の重みを調節する．誤差が出力層側から入力層側へと後向きにさかのぼるためその名が付いた．

ある時刻 t における入力信号 p に対する出力の絶対誤差を式(1)で定義する．

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

ここで，T_{pi} は入力信号 p に対する i 番目出力層のための教師信号，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．この誤差を式(2)にしめすように 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．

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

この誤差が最小となるように，式(3)を用いて結合の重みを調節する．

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

ここで，W_{ij}^k は k 層 i 番目ニューロンと k-1 層 j 番目ニューロンの結合の重みを示す．また，パラメータ α は誤差の振動を減衰させ，解の安定に効果があるが，学習速度の低下や解が局所解に滞留する恐れがある．

そこで，BP 法を拡張した仮想インピーダンス法を用いて，より高速な学習を行う．この方法では式(4)にパラメータ β を含む項が追加された式(4)により重みを調節する．

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

パラメータ β は学習中に強制振動を与えることにより局所的な誤差の極小値（ローカルミニマム）から脱出させる効果がある．

また，成長側抑制学習は，仮想インピーダンス法をさらに拡張した学習方式である．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，ニューロン間の結合を抽出することが困難である．しかし，成長側抑制学習では重みの成長に抑制をかけて重要な結合だけが生き残るように調節するため，ニューロン間の結合が明確になる．成長側抑制学習では，式(4)に成長側抑制項 S が追加された式(5)により重みを調節する．

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

式(5)の成長側抑制項 S は次式で定義される．

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

s：成長側抑制係数，m：k-1 層のユニット数

sgn (x)：x<0 のとき-1，x=0 のとき 0，

x>0 のとき+1 となる関数

ここで，パラメータ s は重みの成長側抑制の効果があり，

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-2　属性の簡素化

		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

s を大きくするほど多少精度は落ちるものの，各ニューロンの結合が簡潔化され，単純な形で規則を得られる．

4.　環境問題への適応 [11)]

4.1　赤潮問題の設定

データマイニングを行う際に必要となるデータベースには表-1 に示す，東京湾の 10 観測ポイントで観測された 15 項目の水質データを用いている．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．東京湾の 10 観測ポイントにおいて 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．なお，全 300 データ中 47 データで赤潮発生が認められた．

そして，ニューラルネットワークを用いて各観測項目と赤潮発生とのネットワーク結合（重みの大きさ）度を求め，この結合度により各観測項目と赤潮発生との関連性を単純な「If ～ then …」形式のルールとして抽出していく．

4.2　ニューラルネットワークによる分析

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各観測項目と赤潮発生との関連性の分析を行う．

(1)　解析モデル・学習データ

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，表-1 の気温～Chl-a の 14 項目を入力値，赤潮発生の有無を出力値に設定して解析を行っている．また，成長抑制係数は予備解析により求めている．

ここでは，入力値を関連性解析の簡素化と各観測値の単位やディメンジョンに統一性がないために，表-2 に示す通り，各観測データの属性を平均値以上･未満のブール属性（0-1）としている．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．また，明らかにノイズとなっているデータは学習中に削除する．

(2)　ネットワークの学習

まず BP 法で学習を行う．ところが，ニューロン間の結合が複雑すぎて判別することが困難である（図-5）．そこで，ある程度までネットワークが構築されたところで，重みの学習法を成長抑制学習に切り替える．その結果，BP 法による学習結果より結合が明確になる（図-6）．しかし，

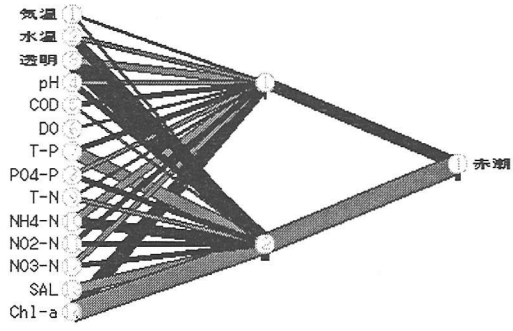

図-5　BP 法による学習結果

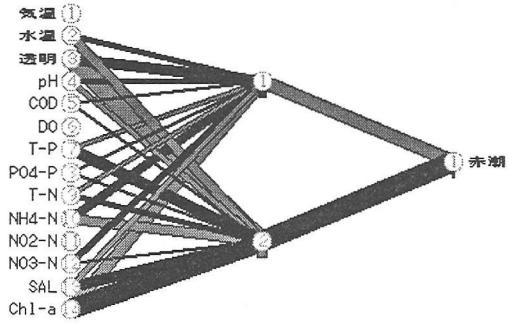

図-6　成長抑制学習の結果

これだけでは単純なルールを抽出できたとはいえない．

図中のネットワークにおいて「＋結合」は赤潮発生と各項目の平均値以上の属性との関連性の強さを，「－結合」は平均値未満の属性との関連性の強さを表している．また，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．

(3) ネットワークの再構築

さらに各ニューロン（観測項目）間の結合を明確にするため，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を削除してネットワークの再構築を行う（図-7）．ネットワークの再構築の結果，重みが小さくなった「気温」，「DO」，「PO-4」，「NO2-N」，「NO2-N」の項目がカットされ，各項目の結合が簡素化されていることが分かる．

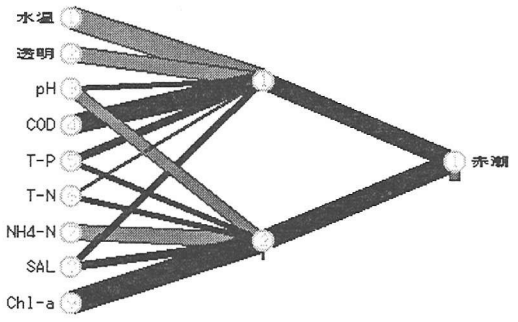

図-7　最構築後の学習結果

(4) 学習結果

「水温」，「透明」，「NH4-N」は平均値未満が，また，「COD」，「T-P」，「T-N」，「SAL」，「Chl-a」は平均値以上が赤潮発生と深く関連しているというルールが得られた．さらに，結合の強さ（太さ）から相関の度合いが評価できる．以上の結果から，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って得られた赤潮発生に関するルールを関連性の大きい順に以下に示す．

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」

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」)

またpHに関しては平均値以下・以上の両方の関連性が抽出されているが，これは他の項目との相互関係が影響していると考えられる．したがって，今回の分析ではルールから除外している．

5. まとめと今後の展望

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用性を検討し，同時に東京湾において実際に観測された水質データと赤潮発生との関連性を解析したところ，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，成長抑制学習を用いて重みの小さくなった結合を削除することにより，ニューラルネットワークの簡素化が実現し，ルール（赤潮発生との関連性）の抽出が容易になった．

(2) 赤潮発生との関連性の強さを基準として，以下に示す3段階に分けられたルールを抽出することができた．
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」

(3) 以上により，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用性が確認された．

また，今後の課題として以下の点が挙げられる．

(1) 実データを用いる場合はデータの前処理が非常に重要である．本研究では数値データを平均値未満・以上のブール属性として考えたが，数値データの前処理手法についてさらなる検討が必要である．また，汎用性が高く，ソフトウェア化された前処理手法の確立が望ましい．

(2) 抽出されたルールに加え，観測ポイント，および，時系列を考慮した解析が必要である．

(3) 本研究ではニューラルネットワークの簡素化のため，中間層を1層（ノード数2）としたが，中間層についてさらなる検討が必要である．

(4) 本来，環境問題には入・出力を分けられないケースが多いため，今後は相互結合型ニューラルネットワークなどによる，入・出力を規定しない解析が必要である

(5) 本来の目的である簡単で汎用性の高いルール抽出を達成するためには，作業全体を通して，システム化されたデータマイニングを確立することが望ましい．

<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，NIKKEI COMPUTER，pp142-147，1996.9.30
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.521-527，1997
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) 中野馨:ニューロコンピュータの基礎, コロナ社, 1990
10) 坂吉則：ニューロコンピューティングの数学的基礎，近代科学社，1995
11) NEUROSIM™/L light，富士通，1996