応用力学論文集 Vol. 3, pp. 99-106 (2000年8月) 土木学会

決定木・GAを用いたデータマイニングによる赤潮発生要因の同定

Data Mining using Decision Tree and Genetic Algorithm For Identification of Red Tide Generation

須藤 敦史[1]・星谷 勝[2]

Atsushi SUTOH and Masaru HOSHIYA

[1]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)
[2]正会員 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 this study consists of the following two topics, one is a basic consideration on a data mining which is the power of current data-processing functions, of interesting knowledge rules from huge database. And the other, we introduce data mining with a view to discuss applications of artificial life and decision tree procedures. Data mining procedures which, decision tree using information entropy theory and genetic algorithm are proposed, and red tide data from Tokyo bay were analyzed. Finally, it is found that the usefulness of these data mining procedures for non-structural system identification.

Key Words: data mining, decision tree, genetic algorithm, information entropy, identification, red tide

1. はじめに

現在,多量のデータベース活用の要求は著しく拡大しており,貴重なデータを効率良く利用するための汎用ツールや関連技術を開発する必要性は高まっているが,一般的に大規模なデータベースは定性・定量データが混在し,相互関係が複雑であるため,現状の解析技術では有効利用がなされていないのが現状である.

そこで,データベースに蓄積されたノイズを含む膨大な生データから,高いレベルで記述された価値ある情報を発掘することを目的としたデータベースからの知識発見(KDD : Knowledge Discovery in Databases)あるいはデータマイニング(DM: Data Mining)[1),2)]の考え方が注目されている.データマイニングとは,膨大なデータの中に存在する隠れた知識(ルール)を客観的に発見することを目標としたものであり,特にマーケティング分野で多くの応用例が報告されている[3),4)].

もっとも,膨大なデータから知識を発見(獲得)しようとする研究は確率・統計,機械学習,データベース技術など多様な枠組みで試られており,探索ツールとしてはファジー理論,決定木(Decision Tree),Genetic Algorithm : GA , Neural Network : NN などの様々な手法が応用されている[5)].しかし現実の問題としてデータベースからの知識獲得には,図-1 に示す概念図のように個別に用いられていた解析技術や手順・手法などを融合した新たなプロセッシング技術あるいは考え方が必要となる.

図-1 データマイニングの概念図

本研究は,条件付き情報エントロピーを用いた決定木の相関探索に対して事象分類における評価方法を提案し,同時に決定木と GA により東京湾で観測された水質調査データと赤潮発生との要因同定を行っている.

2. データマイニング

(1) データマイニングの概要

データマイニングは膨大なデータ中の隠れた知識や規則を客観的に発見することを目的として確率・統計,機械学習,人工知能やデータベース技術を融合してシステム化されたデータ解析技術に対する新しい考え方であり,

この背景には以下の要因が影響していると考えられる.

(a)データベースの発展

コンピュータ技術の進歩で膨大でかつ多様なデータの蓄積が進んでおり,これら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,データベース技術など,従来の解析理論もしくは技術を統合・システム化した新しいデータ解析手法が求められている.

(c)技術のソフトウェア化

実際のデータ解析に適用するため汎用化され,かつ操作手順が簡単なツールが求められている.

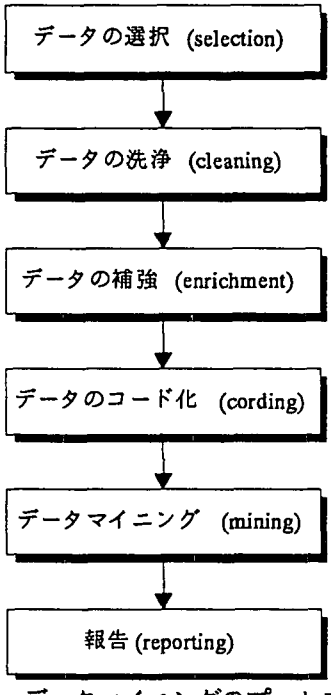

図-2 データマイニングのプロセス

(2) データマイニングの手順

データマイニングの6段階プロセスを図-2に示す.

(a)データの選択(selection)

事象と解析目標を設定し,基本データベースから必要なデータを選定してマイニング用のデータを構築する.

(b)データの洗浄(cleaning)

構築したデータベースから,ノイズや異常値を除去してクリーンなデータを構築する. ただし,この洗浄は事前に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.

(c)データの補強(enrichment)

有用なデータや新たなデータを追加し,他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)の発見を行うものであり,最も重要なプロセスである.

(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,前者は仮説をデータにより検証することを目的とし,後者はデータを単純な表現形式に変換して隠れていたルール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データ解析手法に比べて「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た解析ツール

データマイニングは,基本的にはデータ解析技術・ツールならば何を用いても構わない. したがって,目的に応じて様々なツールを単独あるいは複数組み合わせて使用することも可能であり,本研究では以下の2つの手法に着目してデータマイニングを実データにより行っている.

(a)デシジョンツリー

データ集合(母集団)を属性ごとに分割し,木の枝のように表現する手法であり「決定木」あるいは「判別木」とも呼ばれる.この手法は複数のルールを同時に表現できるため,集合全体を把握するのに有効な手法である.

一方,情報エントロピー[6]は情報が与えられることによる事象の不確定の減少度合いを表しているため,解析する事象間の相関強さを定量的に評価できる利点がある. 加えて相互情報量は条件付き確率と解釈できるため,事象の時間的な前後関係を明確にする可能性を有している.

そこで,本研究では決定木における各事象(属性)の相関関係の評価指標としての情報エントロピーや相互情報量および相関関係の整理を行っている.

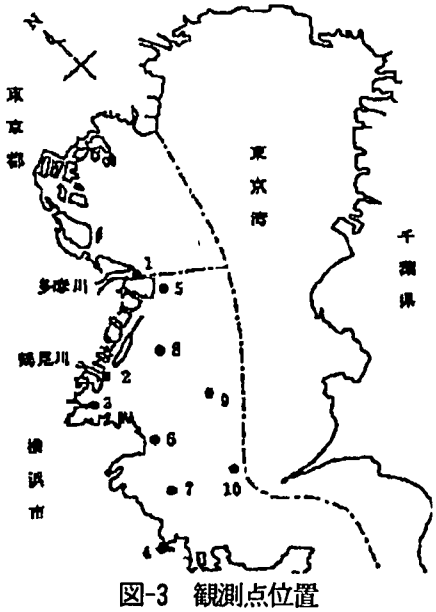

図-3 観測点位置

(b)遺伝的アルゴリズム[7]

生物の進化の過程・集団遺伝のメカニズムを工学的に

モデル化した手法であり,主に組み合せ最適化問題などに応用されている.本研究では属性データのランダムな組み合せを効率よく探索するために利用している.

表-1 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 データのブール属性化

		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

3.東京湾における赤潮と観測項目との相関関係

(1) 水質観測項目

データマイニングに用いる観測データは図-3 に示す東京湾(西側)の10ポイントで観測された水質調査データと赤潮発生の有無を用いており,項目の詳細を表-1 に示す.観測データは1988~92年の5年間における4~9月に観測されたものであり,データ総数は300である.

ここで観測項目のA~Nは数値属性,赤潮に関する項目Oはブール属性(0-1)のデータであるが,マイニングを簡略化するため表-2 に示すように全データをブール属性化し,それらを条件とした「If ~then …」形式のルールを抽出して赤潮発生と観測項目との相関を同定する.

(2) 決定木(デシジョンツリー)による分析

(a) 情報エントロピー

情報エントロピー[6]は情報分野においてC.E.Shannonによって導入され,「現象や情報の不確定の度合い」を表す尺度として用いられているため,事象間の相関関係の強さを評価することが可能となる.

いま,離散型確率分布を有する事象X式(1)を考える.

$$X=\begin{pmatrix} X_1 & X_2 & \cdots & X_m \\ p_1 & p_2 & \cdots & p_m \end{pmatrix} \tag{1}$$

ただし, $0\le p_i\le 1$ かつ $\sum p_i=1$ である.

このとき, Xのエントロピーは次式で与えられ,事象の生起確率で極めて抽象的なものを定量的に評価する指標が'情報エントロピー'である.

$$H(X)\equiv H(p_1,\cdots,p_m)=-\sum_{i=1}^{m} p_i \log p_i \tag{2}$$

一般に,式(2)のlogの底は自然対数e,常用対数10,あるいは2が用いるが,本研究では自然対数eと用いている.

ここで式(1)に示す事象Xとは別に式(3)に示す確率分布を有する事象Yを考える.

$$Y=\begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \\ q_1 & q_2 & \cdots & q_n \end{pmatrix} \tag{3}$$

ただし, $0\le q_j\le 1$ かつ $\sum q_j=1$ である.

また, Y_jが条件として与えられたときのX_iの条件付き確率は,ベイズの定理より次式で与えられる.

$$p_{i|j}=\frac{\pi_{ij}}{q_j}\quad (i=1,\cdots,m \quad j=1,\cdots,n) \tag{4}$$

π_{ij}はX_iとY_jが同時に生起する確率である.このとき,Xの条件付きエントロピーは次式で与えられる.

$$H(X|Y_j)=-\sum_{i=1}^{m} p_{i|j} \log p_{i|j} \tag{5}$$

$$H(X|Y)=\sum_{j=1}^{n} q_j H(X|Y_j) \tag{6}$$

式(6)に示す条件付きエントロピーの範囲は以下となる.

$$0\le H(X|Y)\le H(X) \tag{7}$$

ここで条件が与えられたことによるエントロピーの減少量は相互情報量と呼ばれ,次式のように定義されている.

$$I(X;Y)=H(X)-H(X|Y) \tag{8}$$

式(7),(8)より相互情報量と事象の相関は式(9)のように整理できる.

つまり相互情報量$I(X;Y)$は事象XとYの相関の強さを表す指標として用いることができる.加えて,相互情報量は条件付き確率であるため,事象の時間的な前後関係を明確にすることが可能となる.

$$0 \le I(X;Y) \le H(X) \qquad (9)$$

独立 ← 相関；弱　　完全相関 → 相関；強

（b）観測項目の条件付き確率と相互情報量

ここで気温 A に着目すると，平均値未満のデータは 148 データ存在し，そのうち 23 データにおいて赤潮が発生している．つまり,気温が平均値未満における条件付き赤潮発生確率は次のように算出できる．

$$P(O_1 \mid A_1) = \frac{23}{148} = 0.1554\cdots \cong 0.155 \qquad (10)$$

同様に,平均値以上のデータは 152 データ存在し，そのうち 24 データにおいて赤潮が発生している．つまり気温が平均値以上の条件付き赤潮発生確率は次式となる．

$$P(O_1 \mid A_2) = \frac{24}{152} = 0.1578\cdots \cong 0.158 \qquad (11)$$

ゆえに，気温 A の平均以上・以下について条件付き赤潮発生確率確率の差は次式となる．

$$\left|P(O_1 \mid A_1) - P(O_1 \mid A_2)\right| = 0.0024\cdots \cong 0.002 \qquad (12)$$

上式より，気温 A は赤潮の発生と相関が弱い（条件付確率の差が小さい）ということがわかる．

ここで同様に他属性に対する条件付確率を表-3 に示す．

表-3　条件付き確率の差

X		$P(O_1\|X_1)$	$P(O_1\|X_2)$	$\lvert P(O_1\|X_1) - P(O_1\|X_2)\rvert$
A	気温	0.155	0.158	0.002
B	水温	0.179	0.138	0.041
C	透明	0.272	0.008	0.265
D	pH	0.055	0.231	0.176
E	COD	0.018	0.338	0.321
F	DO	0.071	0.265	0.194
G	T-P	0.096	0.259	0.163
H	PO4-P	0.188	0.092	0.096
I	T-N	0.130	0.210	0.080
J	NH4-N	0.204	0.061	0.143
K	NO2-N	0.175	0.117	0.058
L	NO3-N	0.193	0.068	0.125
M	SAL	0.140	0.165	0.025
N	Chl-a	0.018	0.581	0.563

次に情報エントロピー（相互情報量）による評価を行う．赤潮 O と気温 A に関する相互情報量の算出手順を示す．

（ｱ）赤潮：O の情報ｴﾝﾄﾛﾋﾟｰ

$$H(O) = -P(O_1)\log P(O_1) - P(O_2)\log P(O_2)$$
$$= -\left(\frac{47}{300}\right)\log\left(\frac{47}{300}\right) - \left(\frac{253}{300}\right)\log\left(\frac{253}{300}\right)$$
$$= 0.4341\cdots \cong 0.434 \qquad (13)$$

（ｲ）気温：A_1（平均値以下）に関する条件付きｴﾝﾄﾛﾋﾟｰ

$$H(O \mid A_1) = -P(O_1 \mid A_1)\log P(O_1 \mid A_1) - P(O_2 \mid A_1)\log P(O_2 \mid A_1)$$
$$= -\left(\frac{23}{148}\right)\log\left(\frac{23}{148}\right) - \left(1-\frac{23}{148}\right)\log\left(1-\frac{23}{148}\right)$$
$$= 0.4319\cdots \cong 0.432 \qquad (14)$$

（ｳ）気温：A_2（平均値以上）に関する条件付きｴﾝﾄﾛﾋﾟｰ

$$H(O \mid A_2) = -P(O_1 \mid A_2)\log P(O_1 \mid A_2) - P(O_2 \mid A_2)\log P(O_2 \mid A_2)$$
$$= -\left(\frac{24}{152}\right)\log\left(\frac{24}{152}\right) - \left(1-\frac{24}{152}\right)\log\left(1-\frac{24}{152}\right)$$
$$= 0.4361\cdots \cong 0.436 \qquad (15)$$

（ｴ）気温：A に関する O の条件付きｴﾝﾄﾛﾋﾟｰ

$$H(O \mid A) = P(A_1)\cdot H(O \mid A_1) + P(A_2)\cdot H(O \mid A_2)$$
$$= \left(\frac{148}{300}\right)H(O \mid A_1) + \left(\frac{152}{300}\right)H(O \mid A_2)$$
$$= 0.4340\cdots \cong 0.434 \qquad (16)$$

（ｵ）気温：A に関する相互情報量

$$I(O;A) = H(O) - H(O \mid A) = 0.000005\cdots \cong 0.000 \qquad (17)$$

ここで，同様に他の属性に対する相互情報量を表-4 に示し,各指標に「相関係数」を加えた比較表を表-5 に示す．

表-4　相互情報量

X		H(O)	H(O|X)	I(O;X)
A	気温	0.434	0.434	0.000
B	水温		0.433	0.002
C	透明		0.349	0.085
D	pH		0.402	0.032
E	COD		0.328	0.106
F	DO		0.399	0.036
G	T-P		0.411	0.023
H	PO4-P		0.426	0.008
I	T-N		0.429	0.005
J	NH4-N		0.414	0.020
K	NO2-N		0.431	0.003
L	NO3-N		0.420	0.014
M	SAL		0.434	0.001
N	Chl-a		0.235	0.199

これらの表より,2つの指標（条件付き確率の差，相互情報量）は同様に事象の相関の強さを示しており,比較表から絶対値で比較するとほぼ同じ傾向を示している．

表-5 より相関の強い属性の順に分割したデシジョンツリー1（図-4）を示す．ここで分割終了条件は「確信度が 0 or 1」もしくは「サポートが 0.15 以下」までとし，図中の各属性（ノード）の数値を以下に示す．

一段目：m（m_1 m_2）

現段階まで条件を満足するデータ数を m,その中で赤潮が発生したデータ数が m_1,未発生のデータ数 m_2 である．

表-5 赤潮との相関関係

X		$\lvert P(O_1\lvert X_1) - P(O_1\lvert X_2)\rvert$	I(O;X)	R(O;X)
A	気温	0.002	0.000	0.023
B	水温	0.041	0.002	0.021
C	透明	0.265	0.085	-0.382
D	pH	0.176	0.032	0.303
E	COD	0.321	0.106	0.603
F	DO	0.194	0.036	0.467
G	T-P	0.163	0.023	0.222
H	PO4-P	0.096	0.008	-0.141
I	T-N	0.080	0.005	0.036
J	NH4-N	0.143	0.020	-0.161
K	NO2-N	0.058	0.003	-0.076
L	NO3-N	0.125	0.014	-0.139
M	SAL	0.025	0.001	0.050
N	Chl-a	0.563	0.199	0.588

二段目：確信度

m_1/m（赤潮が発生しているデータ数/条件を満足するデータ数）で定義する指標で条件付き赤潮発生確率である.

三段目：サポート

m/300（条件を満足しているデータ数/データ総数）で定義する指標で現段階までの条件を満足する確率である.

四段目：全赤潮発生数に対する割合

$m_1/47$（現段階の赤潮発生数 m_1/データ総数中の全赤潮発生数（47））で定義する指標であり,本研究でルール評価のための指標として新たに定義するものである.

（c）デシジョンツリーによる解析結果

従来のデータマイニングでは,確信度の高いルールほど評価は高い.ここでデシジョンツリー1（図-4）において矢印が太くなっている関係に着目するとノードの分岐が進むに従い確信度は増加傾向にあるが,全赤潮発生数に対する割合は逆に減少している. 特に「DO：大」という条件が追加されるとその割合の減少は大きいため,条件を満足するデータ数を分母とする確信度の客観性に対して疑問が残る.

そこで,確信度と新たに定義した全赤潮発生数に対する割合が大きい「透明：小」までを示すと**「Chl-a：大 ∩ COD：大 ∩ 透明：小」ならば「赤潮発生」**となる.これは全赤潮発生数に対する割合は 0.872 (41/47) と高く,かつ確信度も 0.651 (41/63) と高い.よって,両指標が高い値を示す属性を赤潮発生の「1次的要因」として考える.

次に確信度のみが高い「T-P：大」までを示すと**「Chl-a：大 ∩ COD：大 ∩ 透明：小 ∩ DO：大 ∩ pH：大 ∩ T-P：大」ならば「赤潮発生」**となる.この要因における確信度は 0.759 (22/29) と高いが, 全赤潮発生数に対する割合に着目すると 0.468 (22/47) と 0.5 を下回った値を示している. つまり赤潮が発生しているデータ（47 データ）の中で半分以上のデータが条件を満足していない. そこで確信度は増加しているが全赤潮発生数に対する割合を減少させてしまう属性については, 赤潮発生に対して「2次的な要因」として考える.よってデシジョンツリー1から導かれる各属性と赤潮発生との相関の強さ（ルール）を評価値から分類すると以下のようになる.

1次的要因：「Chl-a：大 ∩ COD：大 ∩ 透明：小」

2次的要因：「DO：大 ∩ pH：大 ∩ T-P：大」

次に, 各属性を条件付き確率の差および相互情報量を参考にして分割したデシジョンツリー2を図-5に示す.

デシジョンツリー2はデシジョンツリー1に比較して, ツリーの構造が簡素にまとまっているため,この分割は属性間の相互関係が複雑なデータに対して, 効率的な分割が行えると考えられる. 図-5から得られる各属性と赤潮発生との相関の強さを同様に評価値から分類すると以下のようになる.

1次的要因：「Chl-a：大 ∩ 透明：小 ∩ COD：大」

2次的要因：「水温：小」

ここで1次的要因はデシジョンツリー1と同様の結果, また2次的要因は「水温：小」という属性が得られる.

したがって, デシジョンツリー1では獲得できなかった「水温：小」という属性を,各属性間の条件付き確率の差もしくは相互情報量を参考にするにより抽出できた.

（2）GAによる分析

デシジョンツリーを作成することの特徴は,条件付き確率の差もしくは相互情報量を指標として,指標の大きさ（相関の強い）順にツリーを形成していく点にある.

しかし,一般的にデータベースは数多くの属性（項目）を有し,かつデータ総数も膨大であるため,デシジョンツリーのように各属性における条件付き確率の差や相互情報量を求めるには多くの計算時間が必要となる.

そこで本節では相関の順序に関係なく属性間の組み合せを検討するためにGAによるデータマイニングを行う.

（a）GAの解析手順

GA における目的関数（評価）を前節と同様に「①確信度」,「②全赤潮発生数に対する割合」とし,この目的関数を最大にする属性の組合せ（ルール）を探索する最適化問題として,赤潮と属性との相関関係を同定する. ここで組み合せる属性の数は3, 4, 5と限定し,制約条件としてサポートが0.15以下の属性の組み合せは削除している. またGAにおけるパラメータは世代数300, 個体数500, 交叉確率0.9, 突然変異確率0.5を設定した.

（b）GAの解析結果

組み合わせる属性の数が3つの場合のマイニング結果を表-6 に示す. ここで「1」と記されているものは抽出された属性であり, A_1～N_2, m, m_1などはデシジョンツリーの場合と同じである.

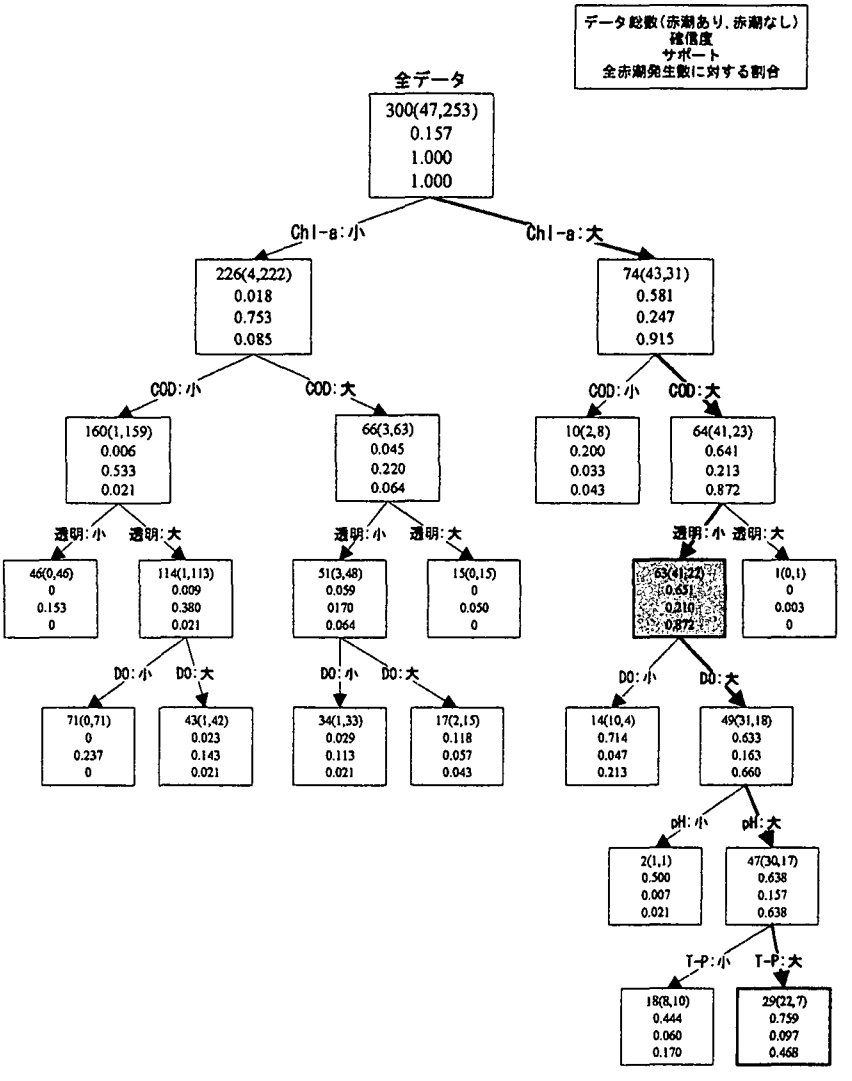

図-4　デシジョンツリー1

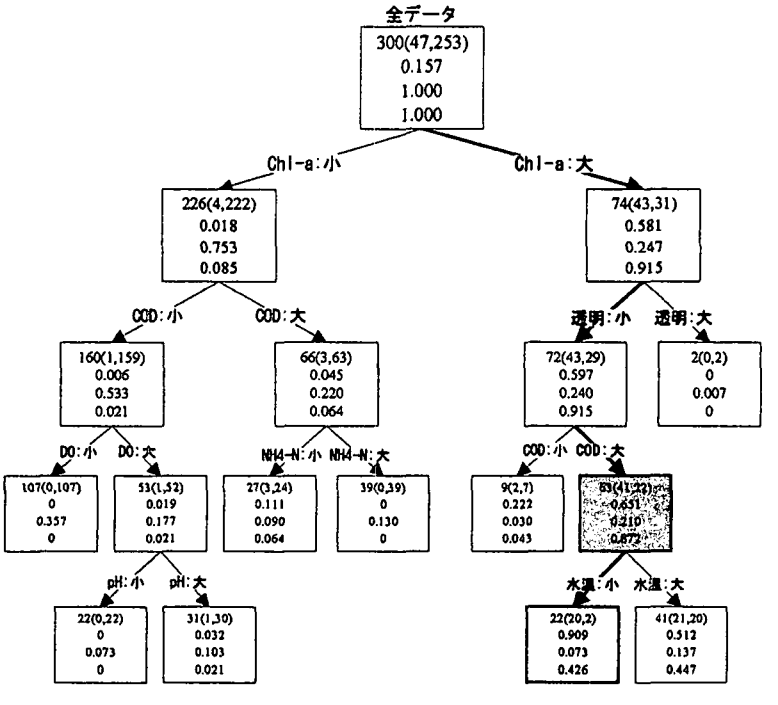

図-5　デシジョンツリー2

表-6 GAによるマイニング結果（属性数：3）

	条件の属性																												m	m_1	確信度	サポート	全赤潮発生数に対する割合
	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	E_1	E_2	F_1	F_2	G_1	G_2	H_1	H_2	I_1	I_2	J_1	J_2	K_1	K_2	L_1	L_2	M_1	M_2	N_1	N_2					
①.「確信度」による評価										1													1					1	51	36	0.70588	0.17000	0.76596
										1									1									1	49	35	0.71429	0.16333	0.74468
										1									1									1	49	35	0.71429	0.16333	0.74468
					1																1							1	49	33	0.67347	0.16333	0.70213
										1									1									1	49	35	0.71429	0.16333	0.74468
										1									1									1	49	35	0.71429	0.16333	0.74468
										1									1									1	49	35	0.71429	0.16333	0.74468
										1									1									1	49	35	0.71429	0.16333	0.74468
					1																1							1	49	33	0.67347	0.16333	0.70213
										1									1									1	49	35	0.71429	0.16333	0.74468
②.「全赤潮発生数に対する割合」による評価					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1					1																		1	63	41	0.65079	0.21000	0.87234
					1			1																				1	64	38	0.59375	0.21333	0.80851
①の合計	0	0	0	0	2	0	0	0	0	8	0	0	0	0	0	0	0	0	7	0	2	0	1	0	0	0	0	10					
②の合計	0	0	0	0	10	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10					
①+②の合計	0	0	0	0	12	0	0	1	0	17	0	0	0	0	0	0	0	0	7	0	2	0	1	0	0	0	0	20					

表-7 GAによるマイニング結果（属性数：4）

	条件の属性																												m	m_1	確信度	サポート	全赤潮発生数に対する割合
	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	E_1	E_2	F_1	F_2	G_1	G_2	H_1	H_2	I_1	I_2	J_1	J_2	K_1	K_2	L_1	L_2	M_1	M_2	N_1	N_2					
①.「確信度」による評価					1					1									1									1	48	35	0.72917	0.16000	0.74468
					1					1									1									1	48	35	0.72917	0.16000	0.74468
					1			1		1																		1	56	36	0.64286	0.18667	0.76596
					1					1									1									1	48	35	0.72917	0.16000	0.74468
					1					1													1			1			50	30	0.60000	0.16667	0.63830
					1					1													1			1			50	36	0.72000	0.16667	0.76596
					1					1									1									1	48	35	0.72917	0.16000	0.74468
					1					1									1									1	48	35	0.72917	0.16000	0.74468
					1					1									1									1	48	35	0.72917	0.16000	0.74468
					1					1									1									1	48	35	0.72917	0.16000	0.74468
②.「全赤潮発生数に対する割合」による評価					1			1		1																		1	56	36	0.64286	0.18667	0.76596
					1			1															1					1	54	36	0.66667	0.18000	0.76596
					1			1		1																		1	56	36	0.64286	0.18667	0.76596
					1			1															1					1	54	36	0.66667	0.18000	0.76596
					1					1													1					1	50	36	0.72000	0.16667	0.76596
					1			1		1																		1	56	36	0.64286	0.18667	0.76596
					1			1		1																		1	56	36	0.64286	0.18667	0.76596
					1					1									1									1	48	35	0.72917	0.16000	0.74468
					1			1		1																		1	56	36	0.64286	0.18667	0.76596
					1					1													1					1	50	36	0.72000	0.16667	0.76596
①の合計	0	0	0	0	10	0	0	1	0	10	0	0	0	0	0	0	0	0	7	0	0	0	2	0	0	2	0	8					
②の合計	0	0	0	0	10	0	0	7	0	8	0	0	0	0	0	0	0	0	1	0	0	0	4	0	0	0	0	10					
①+②の合計	0	0	0	0	20	0	0	8	0	18	0	0	0	0	0	0	0	0	8	0	0	0	6	0	0	2	0	18					

表-8 GAによるマイニング結果（属性数：5）

	条件の属性																												m	m_1	確信度	サポート	全赤潮発生数に対する割合
	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2	E_1	E_2	F_1	F_2	G_1	G_2	H_1	H_2	I_1	I_2	J_1	J_2	K_1	K_2	L_1	L_2	M_1	M_2	N_1	N_2					
①.「確信度」による評価					1			1		1													1					1	47	34	0.72340	0.15667	0.72340
					1			1		1									1				1						53	32	0.60377	0.17667	0.68085
					1					1									1		1		1						50	30	0.60000	0.16667	0.63830
					1					1									1		1		1						50	30	0.60000	0.16667	0.63830
					1					1		1							1				1						47	28	0.59574	0.15667	0.59574
					1			1		1									1				1						53	32	0.60377	0.17667	0.68085
					1					1									1		1		1						50	30	0.60000	0.16667	0.63830
					1			1		1													1					1	47	34	0.72340	0.15667	0.72340
					1			1		1									1				1						53	32	0.60377	0.17667	0.68085
					1			1													1		1			1			57	25	0.43860	0.19000	0.53191
②.「全赤潮発生数に対する割合」による評価					1			1		1									1									1	46	33	0.71739	0.15333	0.70213
					1										1				1				1					1	48	31	0.64583	0.16000	0.65957
					1			1							1				1				1						73	31	0.42466	0.24333	0.65957
					1			1											1		1		1						70	30	0.42857	0.23333	0.63830
					1			1							1				1				1						73	31	0.42466	0.24333	0.65957
					1			1											1				1					1	49	33	0.67347	0.16333	0.70213
								1							1				1				1					1	48	30	0.62500	0.16000	0.63830
					1			1											1				1					1	49	33	0.67347	0.16333	0.70213
					1			1		1													1					1	47	34	0.72340	0.15667	0.72340
					1			1											1				1					1	49	33	0.67347	0.16333	0.70213
①の合計	0	0	0	0	10	0	0	6	0	9	0	1	0	0	0	0	0	0	7	0	4	0	10	0	0	1	0	2					
②の合計	0	0	0	0	9	0	0	9	0	2	0	0	0	0	4	0	0	0	9	0	1	0	9	0	0	0	0	7					
①+②の合計	0	0	0	0	19	0	0	15	0	11	0	1	0	0	4	0	0	0	16	0	5	0	19	0	0	1	0	9					

また,属性数が4つの場合の結果を表-7,5つの場合の結果を表-8に示す.

以下にGAによるマイニング結果で各目的関数の最も高い10個の組み合せを挙げる. ここで( )内の数値は(確信度, 全赤潮発生数に対する割合)を示している.

(ア) 3属性-目的関数①(0.71429, 0.74468)「COD:大 ∩ NH4-N:小 ∩ Chl-a:大」

(イ) 4属性-目的関数①(0.72917, 0.74468)「COD:大 ∩ NH4-N:小 ∩ Chl-a:大 ∩ 透明:小」

(ウ) 5属性-目的関数①(0.72340, 0.72340)「透明:小 ∩ COD:大 ∩ Chl-a:大 ∩ NO3-N:小 ∩ pH:大」

(エ) 3属性-目的関数②(0.65079, 0.87234)「透明:小 ∩ COD:大 ∩ Chl-a:大」

(オ) 4属性-目的関数②(0.72000, 0.76596)「COD:大 ∩ pH:大 ∩ Chl-a:大 ∩ 透明:小」

(カ) 5属性-目的関数②(0.72340, 0.72340)「COD:大 ∩ Chl-a:大 ∩ 透明:小 ∩ NO3-N:小 ∩ pH:大」

以上の結果より「確信度」,「全赤潮発生数に対する割合」のどちらを目的関数に用いてもほぼ同じ属性が選定され,下線が異なる目的関数により得られた属性である.

ここで目的関数①,②により得られた属性を以下に示す.

①,②共通:「Chl-a:大 ∩ COD:大 ∩ 透明:小」

②:「pH:大 ∩ NO3-N:小」, ①:「NH4-N:小」

(3) 専門知識による検証

データマイニングの基本思想は,専門知識は全く参照せずに生データの客観的な解析を重要視することである. この考え方は解析の汎用性と生データが専門家によって恣意的に加工され貴重な情報が失われる危険性や固定観念によりデータの正確な分析ができない事を危惧しているためである. したがって,本研究においてもマイニングの基本思想に従って,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データのみの解析から赤潮発生と水質調査データとの因果関係の解析を試みている.

ここでは,あえて現在まで赤潮と水質調査データに関して得られている専門的な知見(東京湾の赤潮に関する調査報告書[8])を参照して,本研究で得られた結果の検証を行う. 調査報告書より赤潮発生時の水中でCOD, Chl-aが大きいのは動植物プランクトンのデトライタスに由来している.加えてT-P, T-Nも大きい値を示すと報告されており,今回の解析でも同様な結果が得られている.

また,赤潮発生と水温,塩分濃度との関連性も最近論じられている.ここでデシジョンツリーでは「水温:小」やGA(表-7)では「SAL:大」という属性が得られており,新たな知識獲得を示唆するものと考えられる. よって,今回のデータマイニングの結果は妥当であると考えられ,GAを利用する事や相互情報量を参考にすることで,データからより詳細な情報の抽出が可能になったと言える.

4. 結 論

本研究ではデータマイニングにおける事象間の相関関係が情報エントロピーの減少量(相互情報量)によって評価できることを示した. 同時に実データによる解析例として東京湾で観測された水質観測データを用いて赤潮発生要因の同定をデシジョンツリーおよびGAを用いて行い, 以下の結論が得られた.

(1) 各属性において条件付き赤潮発生確率を算出した結果, 赤潮発生と相関の強い属性として「Chl-a」,「COD」,「透明」が抽出された. また相互情報量を用いて評価した場合も同様の結果が得られた.

(2) デシジョンツリーにより属性間の相互関係を詳しく分析し, 2次的な要因として「DO:大 ∩ pH:大 ∩ T-P:大」や「水温:小」という属性が赤潮発生に対して相関を有する結果となった.

(3) GAによる分析からも「Chl-a:大 ∩ COD:大 ∩ 透明:小」の属性が選択された. しかし,デシジョンツリーでは得られなかった「pH:大 ∩ NO3-N:小」,「NH4-N:小」が要因として抽出された.

また, 今後のデータマイニングの課題としては以下の点が挙げられる.

(1) データの前処理が重要となる. 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たが, 今後は数値データの前処理について検討が望まれる.

(2) 本研究では時間的な因果関係までは分析できなかったが,今後は時間的な分析を行い,「A→B→赤潮発生」という因果関係の同定が期待される.

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) 大規模データベースからの知識獲得, 人工知能学会誌, Vol.12,No.4,pp496-549, 1997.7

3) 徳山豪: データマイニングに使われる最適化の数理, 応用数理, VOL.6,NO.4,pp303-313, 1996.12

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, 1996.9.30.

5) 須藤敦史, 高須光朗, 星谷勝: 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, 応用力学論文集, Vol.2,pp.83-90,1999.

6) 有本卓: 確率・情報・エントロピー, 森北出版, 1992.

7) 北野宏明: 遺伝的アルゴリズム, 産業図書, 1993.

8) 二宮勝幸: 横浜市沿岸および沖合域の水質変動特性, 横浜市環境科学研究所資料, 東京湾の富栄養化に関する調査報告,No.117,pp.9-26,1995.

(2000年4月21日 受付)