データマイニングにおけるGA・情報エントロピーの応用

Data Mining using Information Entropy and Genetic Algorithm

須藤 敦史* ・ 渋谷 卓**・ 星谷 勝***
Atsushi SUTOH , Taku SHIBUYA and Masaru HOSHIYA

* 博士(工学)（株）地崎工業 土木技術部 主席研究員(〒105-8488 東京都港区西新橋2-23-1)
** （株）地崎工業 情報システム部 (〒105-8488 東京都港区西新橋2-23-1)
*** 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 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*

1．はじめに

現在,データウエアハウス活用の要求が多くなっているが定性・定量的データが混在し,加えて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.そこで価値ある情報や知識の発掘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]が注目されている.

もっとも,膨大なデータから知識を得ようとする研究は確率・統計,機械学習など多様な枠組みで試みられている[3]が，データマイニングは図-1 に示すようにデータ中の隠れたルールを客観的に発見するための現実問題として従来の手法を統合した新たなプロセッシング技術あるいは考え方であり,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[4],[5]．

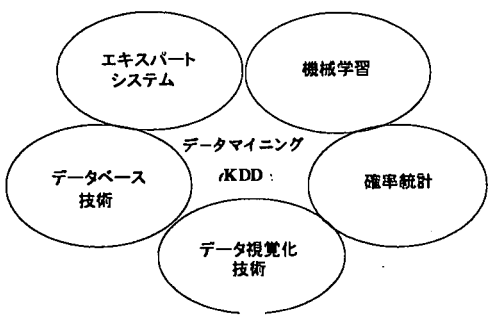

図-1 データマイニングの概念図

本研究は,事象の相関関係（決定木）に条件付き情報エントロピーを用いた評価方法を提案し,同時に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相関解析を行っている.

2．データマイニング

データマイニング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

従来のデータ解析理論もしくは技術を統合・システム化した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

実データを解析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.

一般的にデータマイニングは「仮説検証型」と「仮説生成型」に分けられ,前者は仮説のデータによる検証すを,後者はデータを単純な表現形式に変換して隠れたルールの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に比べて「状況予測」より「結果解釈」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

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が目的に応じて様々なツールを単独あるいは複数組合せて使用する.本研究では以下の２手法により

データマイニングを行っている.

(a)デシジョンツリー

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,事象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[1]は事象の不確定の度合いを表しているため,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もを有している.

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.

(b)遺伝的アルゴリズム[1]

生物の進化過程・集団遺伝のメカニズムを工学モデル適用にした手法であり,主に組み合せ最適化問題などに応用されている.本研究では事象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.

4.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

観測データは図-2 に示す東京湾(西側)の 10 点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.

ここで観測データは 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.

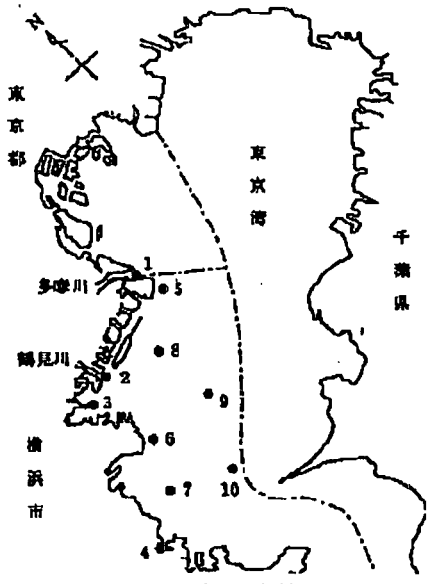

図-2 観測点位置

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,簡略化するため表-2 のように全データを平均値以上・以下のブール属性化し,それらを条件とした「If ~then …」形式のルールとして赤潮発生と観測項目との相関解析を行う.

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

(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[1]は「現象や情報の不確定の度合い」を表す尺度でり,事象間の相関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 を考える.

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

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

このとき事象 x の情報エントロピーは式(2)となる.

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

ここで式(3)に示す確率分布を有する事象 y を考える．

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

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

また Y_j が条件として与えられたときの X_i の条件付きエントロピーは式(4)で与えられ，範囲は式(5)となる．

$$H(X|Y)=\sum_{j=1}^{n} q_j H(X|Y_j) \qquad (4)$$

$$0 \le H(X|Y) \le H(X) \qquad (5)$$

ここで条件が与えられたことによる条件付き情報エントロピーの差は相互情報量と呼ばれ，式(6)となる．

$$I(X;Y)=H(X)-H(X|Y) \qquad (6)$$

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，加えて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．

$$\underbrace{0 \le I(X;Y)}_{\text{独立}} \le \underbrace{H(X)}_{\text{完全相関}} \qquad (7)$$

相関；弱　←　　→　相関；強

（b）事象（観測項目）の相互情報量

各事象に対する条件付き確率の差，相互情報量に相関係数を加えた比較表を表-3に示す．

表-3　赤潮との相間関係

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

表より，条件付き確率の差，相互情報量は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．

ここで表-3 より相関の強い順に分割した決定木1を図-3 に示す．決定木における分割終了条件は「確信度が 0 or 1」もしくは「サポートが 0.15 以下」まで，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$_2$)　条件を満足する総数：m，その中で赤潮発生数：m$_1$，未発生数：m$_2$

二段目：確信度　m$_1$/m（赤潮の発生数/条件を満足する数）で定義される指標で条件付き赤潮発生確率

三段目：サポート　m/300（条件を満足する数/データ総数）で定義される指標で現段階まで条件を満足する確率

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$_1$/総数中の全赤潮発生数（47））詳細な相関関係評価のための指標として新たに定義

（c）決定木による解析結果

従来では，確信度が高いほどルールとしての評価は高いが，決定木1（図-3）において矢印が太い関係に着目するとノード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」という条件が追加するとその減少が大きいため，条件を満足するデータを分母とする確信度の客観性に疑問が残る．

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事象までを示すと「**Chl-a：大 ∩ COD：大 ∩ 透明：小**」ならば「**赤潮発生**」となり，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事象までを「1次的要因」として考える．

次に確信度のみが高い事象は「**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．この確信度は 0.759 (22/29) と高いが，全赤潮発生数に対する割合は 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 47 ケースの中で半分以上のデータがこの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が減少してしまう「2次的な要因」として考える．

よって決定木1から導かれる相関の強さ（ルール）を上記の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，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決定木2を図-4に示す．

決定木2ではツリーの構造が簡素化されているため，事象の相互関係が複雑なものに対して，効率的な分割が行えると考えられる．

ここで図-4 から得られる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」

1次的要因は決定木1と同様の結果，また2次的要因は「水温：小」という事象が追加されてる．

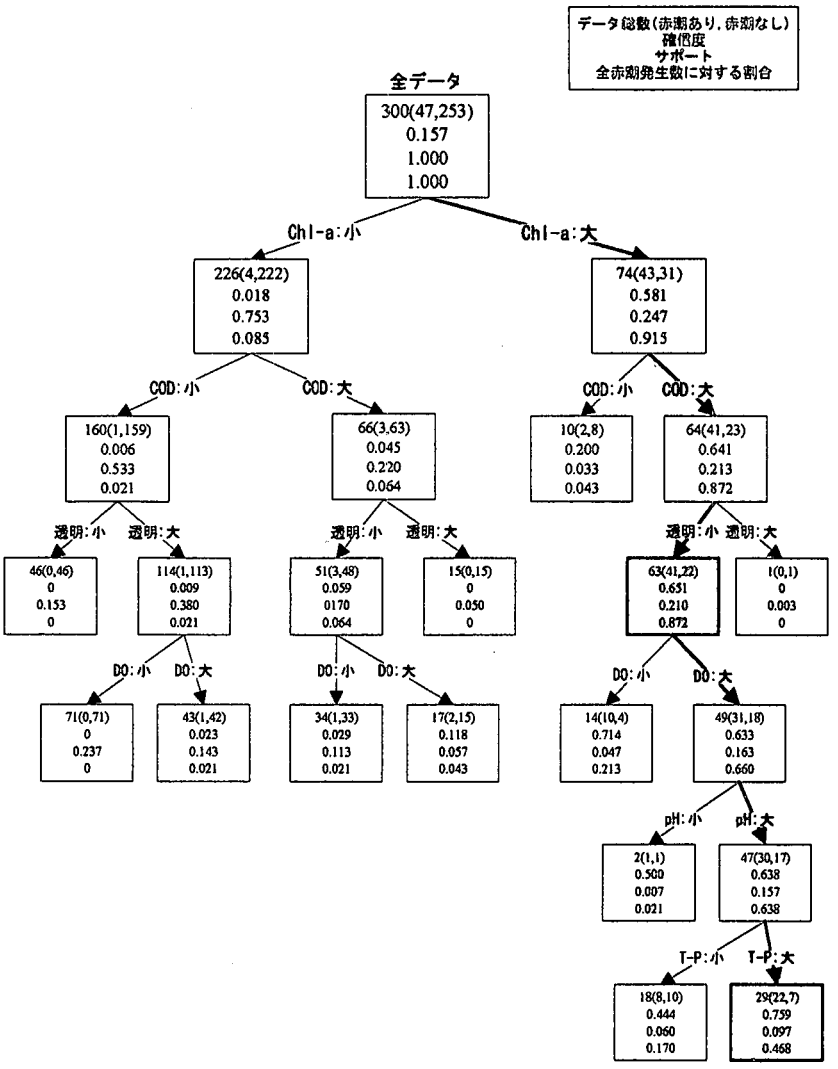

図-3 決定木1

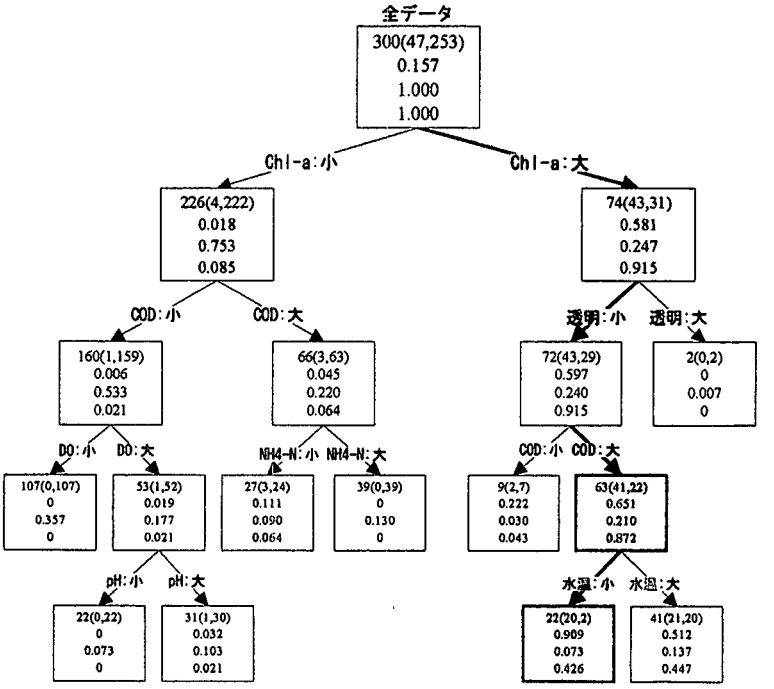

図-4 決定木2

したがって,決定木1では獲得できなかった「水温：小」が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

決定木の特徴は,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）順に枝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事象（項目）を有し,データ総数も膨大であるため,各事象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.

そこで相関の順序に関係なく事象の組合せを検討するためにGAを用いたよる相関の強さ（ルール）解析を行う.

（a）GAの解析手順

GAにおける目的関数を「①確信度」,「②全赤潮発生数に対する割合」として,これを最大にする事象の組み合わせを探索する最適化問題として,赤潮と事象との相関関係（ルール）を求める.

ここで組み合わせる属性の数は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

組み合わせる事象数が3つの結果を表-4に示す．表中の「1」は選択された事象であり，A_1〜N_2，m，m_1は決定木と同様である．また,事象数が4つの結果を表-5,5つの結果を表-6に示す.

以下に各目的関数の高いの組み合せを挙げる．ここで（）内は確信度，全赤潮発生数に対する割合である.

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

(イ) 4事象-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事象-目的関数②（0.65079，0.87234）「透明：小 ∩ COD：大 ∩ Chl-a：大」

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

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

以上より「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事象が選定される.ここで,下線はどちらかの目的関数により得られた属性である.

また目的関数①,②により得られた事象を以下に示す.

①,②共通：「Chl-a：大 ∩ COD：大 ∩ 透明：小」

②：「pH：大 ∩ NO3-N：小」，①：「NH4-N：小」

5．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条件付き確率の差および相互情報量によって評価できることを示し,同時に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相関解析を条件付き確率の差および相互情報量を用いた決定木の分類および膨大な組み合わせのためにGAを用いて行い，以下の結論が得られた.

(1) 条件付き確率では「Chl-a」，「COD」，「透明」の事象が抽出さ,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) 条件中の割合の評価により,2次的な要因「DO：大 ∩ pH：大 ∩ T-P：大」や「水温：小」が赤潮発生に対して相関を有する結果となった.

(3) GAによる解析からも「Chl-a：大 ∩ COD：大 ∩ 透明：小」が選択された．また決定木では得られなかった「pH：大 ∩ NO3-N：小」，「NH4-N：小」が選択された.

また,今後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ため,効率的な数値データの前処理について検討が望まれる.

(2) 「A→B→赤潮発生」という因果関係などの時間的な分析が期待される.

参考文献

1) Pieter Adriaans, 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.

表-4　事象：3の結果（GA）

	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-5　事象：4の結果（GA）

	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-6　事象：5の結果（GA）

	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					