データマイニングにおけるGA・情報エントロピーの適応について

○㈱地崎工業　　正会員　須藤　敦史
武蔵工業大学　　正会員　星谷　勝

1．はじめに

コンピュータ技術の進歩によりデータウエアハウス利用要求が高まっているが定性・定量データ間の関係が複雑であるため,現状の解析技術では活用されていない.そこで価値ある知識の発掘を目的としたデータマイニング(DM:Data Mining)[1),2)]が注目され,統計解析をはじめとして決定木・ニューラルネットワーク:NN・遺伝的アルゴリズム:GA・ファジー理論など様々な学習ツールが適用されている.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手法を示し,同時にGAのデータマイニングへの適用を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通して行っている.

2．赤潮と観測データ

観測データは**表-1** 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1988～92年(5年間)4～9月に観測したものでデータ総数は300である.ここで観測項目A～Nは数値属性,赤潮に関する項目Oはブール属性（0-1）であるが,簡略化のため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 …」ルールで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　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

3．デシジョンツリー

母集団を属性ごとに分割して木の枝のように表わして複数のルールを同時に表現するため,事象全体を把握するのに有効である.一方,情報エントロピー[3)]は「現象や情報の不確定の度合い」を示す尺度であり,事象間の相関強さを定量的に評価する利点がある.いま Y_j が与えられたとき X_i の条件付きエントロピーは式(1),範囲は式(2)となる.

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

$$0 \le H(X|Y) \le H(X) \qquad (2)$$

ここで相互情報量 $I(X;Y)$ (条件付き情報エントロピーの差)は事象 X と Y の相関の強さを表す指標となる.

$$I(X;Y)=H(X)-H(X|Y) \qquad (3)$$

$$0 \le I(X;Y) \le H(X) \qquad (4)$$

表-1 より,相互情報量(相関関係)の大きい(強い)順に分割した決定木を**図-2** に示す．ここで終了条件は「確信度 m_1/m（赤潮の発生数/条件を満足する数）が0 or 1」もしくは「サポート m/300（条件を満足する数/データ総数）が0.15以下」

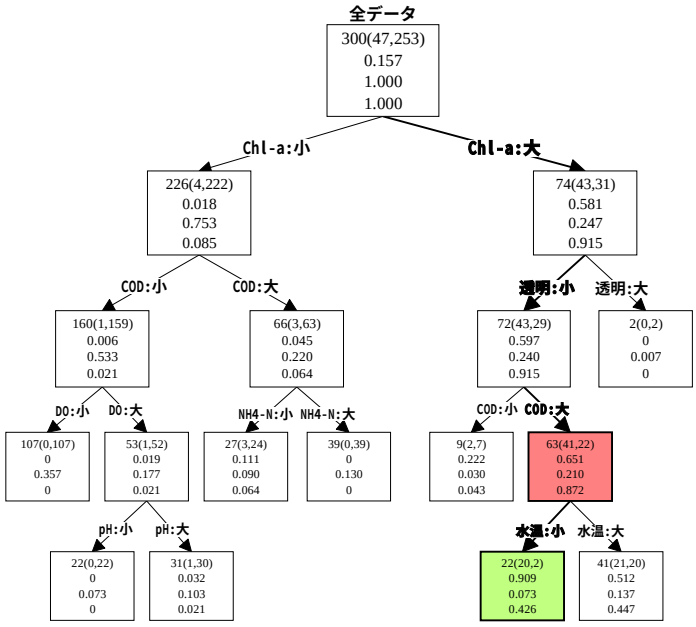

図-1　決定木

キーワード:　知的情報処理，人工生命技術，データマイニング，　情報エントロピー　，環境モニタリング
連絡先（〒105-8488　東京都港区西新橋2-23-1　TEL 03-3592-6955　FAX 03-3502-2646　E-mail 1714@chizaki.co.jp）

表-3　GAの結果（事象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					

ここで条件を満足する総数 m,赤潮発生数 m_1 未発生数 m_2 である.

以上,相互情報量を参考にして分割した決定木ではツリー構造が簡素化されるため,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．ここで**図-1** 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示すと１**次的要因**:「**Chl-a**：大 ∩ **透明**：小 ∩ **COD**：大」，２**次的要因**:「**水温**：小」となった.

4. 遺伝的アルゴリズム [3)]

決定木は相互情報量の大きさ（相関の強い）順に枝を形成していくが,一般的にデータベースは数多くの事象（項目）を有し,かつデータ総数も膨大であるため,全組み合わせの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み合わせを効率よく探索するために GA を用いたよる相関の強さの解析(ルールの抽出)を行う.

ここで 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,赤潮と事象との相関関係（ルール）を求める.組み合わせる事象(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 以下の組み合わせは削除している．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.

組み合わせた事象数が３つの場合の結果を**表-3** 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ている.

ここで目的関数①,②により得られた事象は①,②**共通**:「**Chl-a**:大 ∩ **COD**:大 ∩ **透明**:小」，②:「**pH**:大 ∩ **NO3-N**：小」，①：「**NH4-N**：小」となった.

4．まとめ

本研究ではデータマイニングおいて相互情報量によって事象間の相関関係が評価できることを示し,同時に膨大な組み合わせを効率よく探索する GA の適用を東京湾で観測された水質観測データを用いて行い 1)「Chl-a:： ∩ 透明：大 ∩ COD：大」や「水温：小」が赤潮発生に対して相関を有する結果を得た.また GA による最適化解析からも同様な事象が選択され,また決定木では得られなかった「pH：大 ∩ NO3-N：小」,「NH4-N：小」も選択された.

また,今後の課題としては 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の検討,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) 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
2) 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
3) 有本卓：確率･情報･エントロピー，森北出版，1992.
4) 北野宏明：遺伝的アルゴリズム，産業図書，1993.