# 情報エントロピーを用いたデータマイニングによる環境データと赤潮との相関解析

## Data Mining using Information Entropy for Association Rules of Red Tide

（株）地崎工業 土木技術部　　正会員　須藤　敦史(Atsushi SUTOH)
（株）地崎工業 情報システム部 ○正会員　渋谷　卓(Taku SHIBUYA)

### 1．はじめに

大規模なデータベースから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．

データベースから知識を得ようとする試みは確率・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現実問題としてのデータプロセッシングの概念もしくは考え方である[1),5)]．

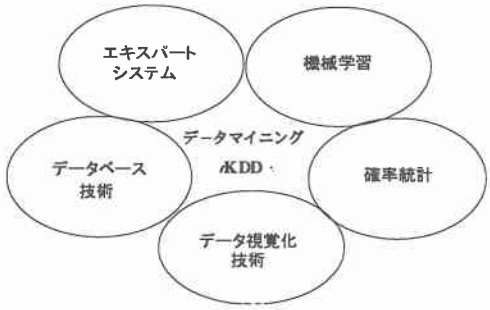



図-1　データマイニングの概念図

本研究では,事象の相関関係を条件付き情報エントロピーを用いた決定木で評価し,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．

(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

従来の解析理論や技術を統合・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の汎用化され,かつ操作手順が簡単なツールが求められている．

データマイニングでは仮説をデータによる検証する「仮説検証型」と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「仮説生成型」に大別されている．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データマイニングの特徴である．

### 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目的に応じて様々なデータ解析技術やツールを単独あるいは複数組合せて使用するが,本研究では決定木(デシジョンツリー)を用いている．

この方法は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,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．

また,情報エントロピー[6)]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．


図-2　観測点位置

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.

## 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.

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:$O_2$

ここで観測データは 1988～92 年(5 年間)の 4～9 月に観測された総数 300 個のデータであり,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.

### （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.

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

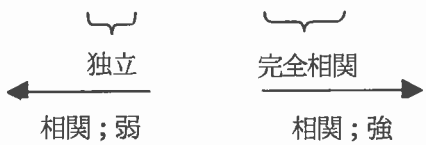
$$0 \le H(X|Y) \le H(X) \qquad (5)$$

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.

$$0 \le I(X;Y) \le H(X) \qquad (7)$$




#### （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. 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,比較表の絶対値からほぼ同じ傾向を示している.

相関の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.

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 |

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したものである．

（c）決定木による解析結果

従来，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の太い矢印に着目するとノード分岐が進むに従い確信度は増加するが，全赤潮発生数に対する割合は減少している．特に「D0：大」では減少が大きいため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は「**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するが全赤潮発生数に対する割合が減少する「2次的な要因」と考える．

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ている．

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ため，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ことにより抽出できている．

## 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」の事象が抽出され，また相互情報量で評価した場合も同様の結果が得られた．

(2)　全赤潮発生数に対する割合の評価により，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)　今後の課題としては「A→B→赤潮発生」のような時間的な因果関係分析が期待される．

### 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.

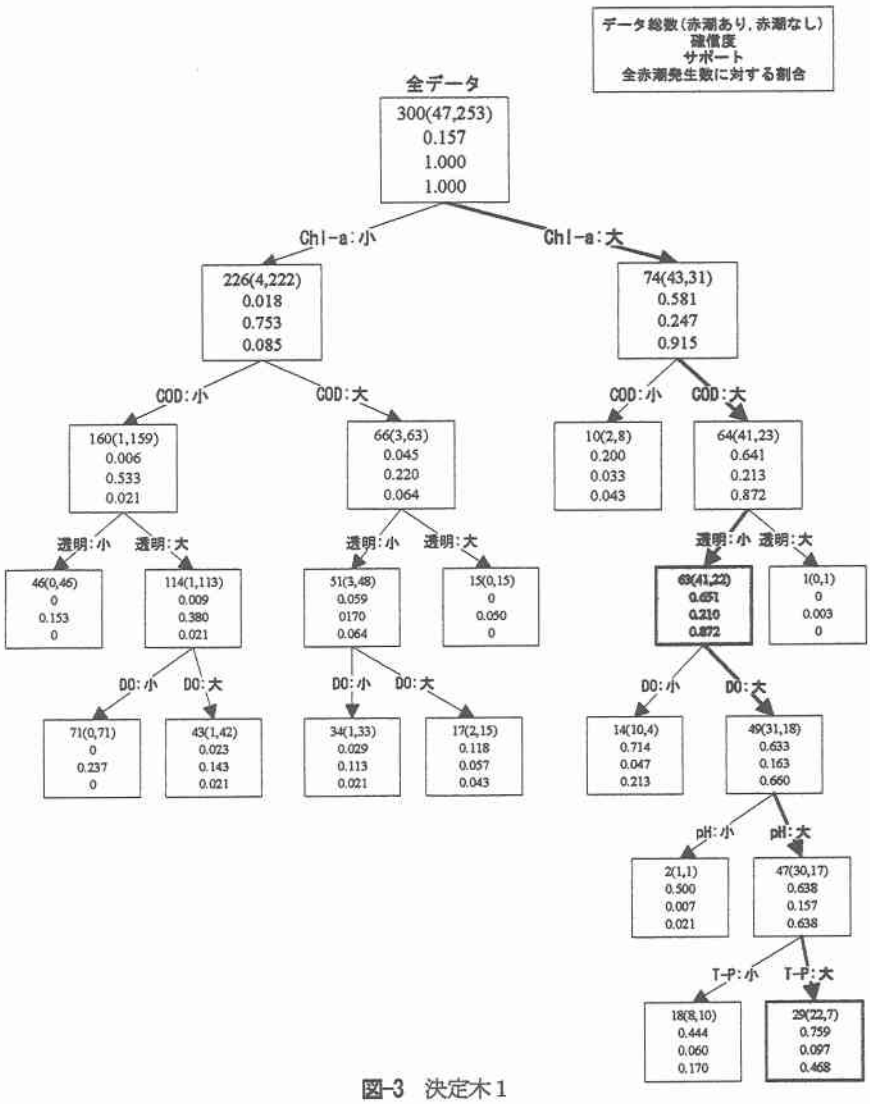



図-3 決定木1

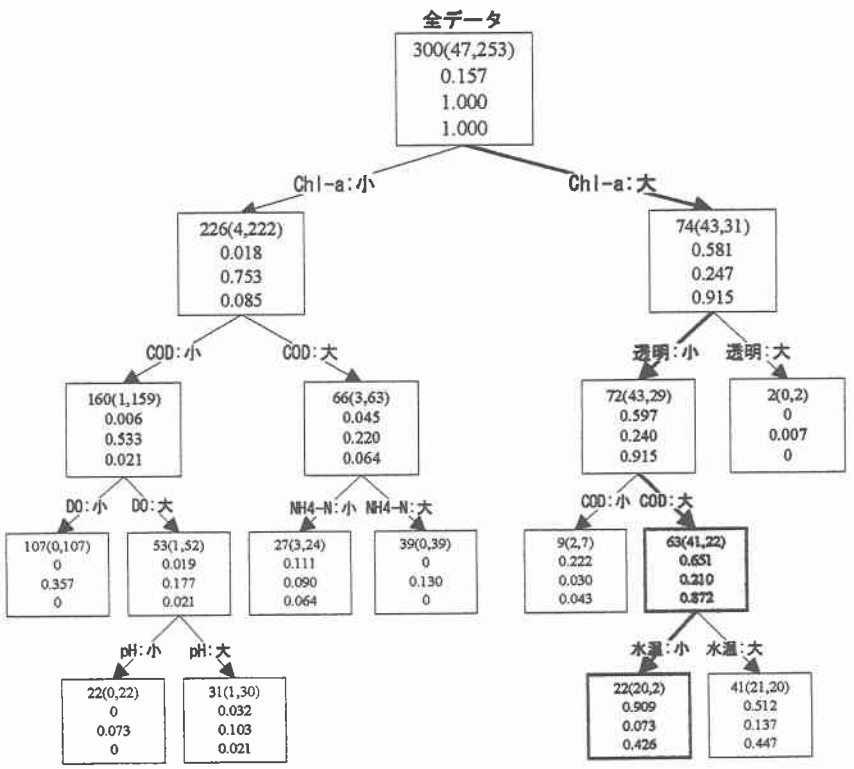



図-4 決定木2