

非構造システムの同定・逆問題における データマイニングによる相関ルール抽出について

A Basic Consideration on Non-structural Inverse or Identification Problems
using Mining Algorithms for Association Rules

須藤 敦史¹・佐藤 大介²・星谷 勝³

Atsushi SUTOH, Daisuke SATO 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 修士(工学) 武蔵工業大学工学部工学研究科 (〒158-8557 東京都世田谷区玉堤1-28-1)

³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 recent years, data accumulation has been remarkably increased, and it is pointed out that it is necessary to restore data in order, and to discover rules hidden behind data. The notion of data mining is first clarified and is interpreted as a non-structural inverse or identification problem. Then, within the scope of data mining, numerical analyses are carried out to find out rules on three sets of data whose features are different, that is, (1) occurrence of a red tide in Tokyo bay, (2) causal relationship of water quality in Lake Shinji and Nakaumi. Finally, a neural network algorithm is found that the usefulness of a data mining procedure for non-structural system identification.

Key Words: data mining, environment data, neural network, inverse problems, identification

1. はじめに

コンピュータ技術の急速な発展により、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、定性・定量データの混在化、形式(フォーマット)の非統一性、相互の複雑性などの理由により、価値のある知識の獲得が非常に難しくなっており、保有する貴重なデータを有効に活用してないのが現状である。

一方、変化の著しい情報化社会では、膨大な情報の中から冗長な部分を整理し、表面上表れない現象間の法則やルールを早急に見つけ出すことが必要となるため、データの背後に隠れた『知識やルールもしくは現象や事象間の因果関係』を客観的に探索するための手法の開発が望まれている。

このような背景のもと、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)}、以下DM)が注目されている。DMとは膨大な生データの中から、高いレベルで記述された価値あるルールや法則を抽出することを目標としたもので、流通・マーケティングや医療の分野で応用例が多数報告されている³⁾。

もっとも、従来のデータ分析あるいは処理においてもデータの中から知識を獲得しようとする試みは、確率・統計、機械学習、データベース技術など様々な研究の枠組みの中で行われており、集合理論、決定木(Decision Tree)、Genetic Algorithm: GA、Neural Network: NN、人工知能(Artificial Life)技術などが探索ツールとして用いられている^{4), 5), 6), 7)}。

しかし、現実問題としてデータベースから新たな知識獲得は容易なものではないため、DMは従来個別に用いられている解析技術や手順・手法などを融合した総合的データプロセッシング技術(考え方)とも言える。

一方、DMは現象・事象の結果(観察)をもとに対象とする現象や事象間の因果関係やその原因を発見もしくは探索・把握することより、逆解析的なアプローチにほかならない。

しかし、DMにおいて最初に与えられるものは現象・事象の観察結果であり、その中から何か新しい知識を把握しようとするものである。さらに現象・事象の関係式や支配方程式がどのように表されるか、それに対する境界条件や各種パラメータの存在の有無までも明確ではなく、加えて入力値が特定できないため、非常に特殊な(究極

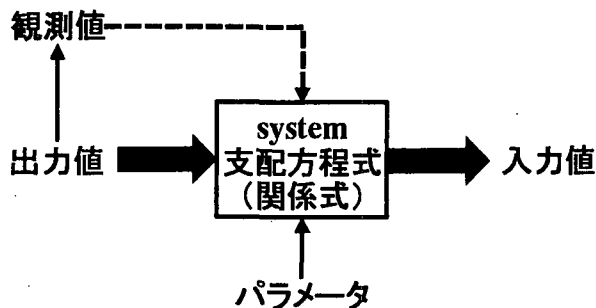


図-1 一般的な逆問題

の) 逆問題とも言える。

そこで、本研究では DM を非構造システムにおける特殊な逆問題として捉え、ニューラルネットワーク (Neural Network, 以下 NN) を探索ツールとして観察事象の解析を数的な処理ではなく集合的な考えで行うことを目的としている。

加えて、NN を用いた DM の実データ解析への適用性を検証するため、

- (1) 東京湾で観測された赤潮発生に関する水質データ、
- (2) 宍道湖・地中海で観測された水質データ、を用いて相関関係の解析を試みている。

また、NN は階層型と相互結合型に分類されるが本研究では適用例の多い階層型ネットワークを採用し、NN 中の重みの大きさからデータ間の相関・関連性の強さを算出している。さらに、階層型による擬似的相互結合ネットワークに対する基礎検討も行っている。

2. データマイニングと同定・逆問題

(1) データマイニング

DM にはデータベース技術からはじまり、データの可視化技術、統計的手法、人工知能、機械学習、各種推論技術、オペレーションリサーチなど非常に広い範囲のコンピュータ技術や数理解析手法を用いられているが、実際の問題としては膨大な生データへDMを適用するためには、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のが現状である。

一般的にデータマイニングは様々な仮説をデータにより検証する「仮説検証型」とデータを単純な表現形式に変換して(a)~(f)に示すデータ間のルールを発見する「仮説生成型」に分類されるが、両者とも従来のデータ解析に比べて「状況の予測」より「結果の解釈」に重点が置かれている。

(a) 相関ルールの発見 (association rules)

データ間に存在する (隠れた) 相関関係の抽出する。

(b) 分類ルールの発見 (classification rules)

データをいくつかのクラスに分類するための関数やルールを把握する。

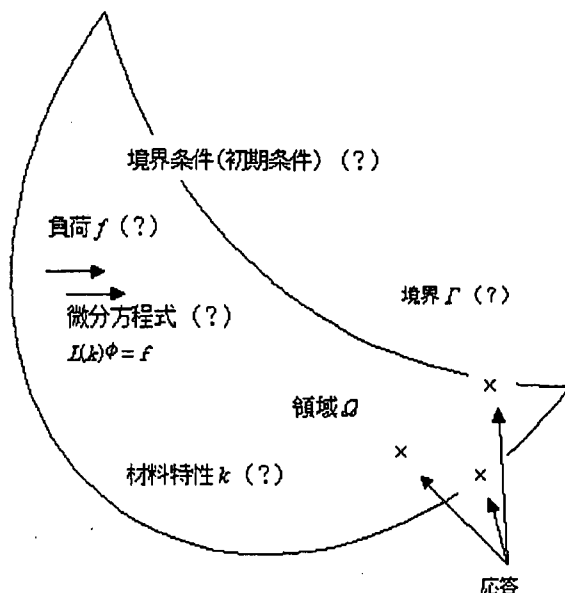


図-2 逆問題

(c) クラスタリング (clustering)

データが属する有限個のカテゴリ集合を同定する。

(d) 回帰分析 (regression)

データ項目と実数の予測変量とを写像づけ、変量間の関係を見いだす。

(e) データ集約 (summarization)

データ集合に対する簡素な記述を見いだす。

(f) 変化の発見 (reporting)

時間的・空間的に変化するデータに対して、もっとも変化の著しい部分を同定することであり、データ変化の順序性の発見も含まれる。

(2) 逆・同定問題

逆・同定問題 (Inverse / Identification Problems) は工学の様々な分野で扱われて (現れて) おり、出力値をもとに入力値やその条件などを求めるあるいは観測値や観察結果から原因を推定するものである。

一般的な逆問題は、図-1 に示すように順問題の手順とは逆に、出力値をもとに未知のパラメータや入力値を同定するものである⁹⁾。

いま、状態量 ϕ の空間変化あるいは時間変化を支配している微分方程式は以下となる。

$$L(k)\phi = f \quad (1)$$

$L(k)$: 材料特性 k を含んだ作用素 (Operator), f : 負荷
ここで、逆問題は図-2 に示すよう (a) から (e) のうちの1つ、あるいは複数を同定しようとするものである。

(a) 境界/領域逆問題: 対象とする領域 Ω とその境界 Γ の位置と形状の支配方程式に対する逆問題: 物理量 ϕ に関する場の支配方程式 ($L\phi = f$)

(b) 境界値/初期値逆問題: ϕ およびその微係数に関

知識の発見・把握

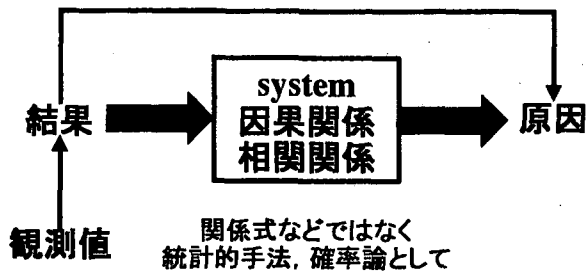


図-3 逆問題としてのデータマイニング

する境界条件 (初期条件も含む)

(c) 負荷逆問題: 負荷 f (入力)

(d) 材料特性逆問題: 支配方程式に含まれる材料特性 (パラメータ) k の分布 (材料特性逆問題)

また, 実際に直面する様々な逆問題を考えてみると, 解決に必要な情報が常にすべて完全な形として与えられていることは稀であり, さらに材料特性・支配方程式やそれに対応した境界条件をどのように表すかは全く分からない場合にもしばしば遭遇する。

上記のような問題は不適切性 (ill-posed) を有するとされ, 逆解析特有の問題であり, 逆問題が適切 (well-posed) であるためには, 以下に示す3つの要件を満足しなければならない。

(1) 存在性 (existence), (2) 一意性 (uniqueness), (3) 連続性 (continuity) あるいは安定性 (stability)

そこで, このような要件を (擬似的に) 満足させて逆問題を解く様々な最適化手法・技法などや考え方を適用もしくは導入しているのが現状である。

(3) データマイニングと同定・逆問題

DM は観測される事象や現象に対して, その相関関係や因果関係を発見・把握することを目的としているため, 観測値から対象とする関係やそのシステムを推定・同定する逆解析的なアプローチと言える。

また, DM は図-3 に示すように上記の一般的な逆問題と異なり, 現象・事象の関係式や支配方程式がどのように表されるか, それに対する境界条件や各種パラメータの存在の有無までも明確ではなく, 入力値が特定できない特殊な問題でもある。

加えて, DM において最初に与えられるものは現象の観察結果のみの場合が多く, つまり観測値をもとに現象を理解する過程は逆問題的アプローチに他ならない。

したがって, DM は従来におけるシステムの関係式・支配方程式やパラメータ・入力の同定問題ではなく, 観察された事象の統計解析あるいは集合論としてのデータプロセッシングとして考えなければならない特殊な (究極の) 逆問題に相当し, 土木工学において本質的な逆問題であ

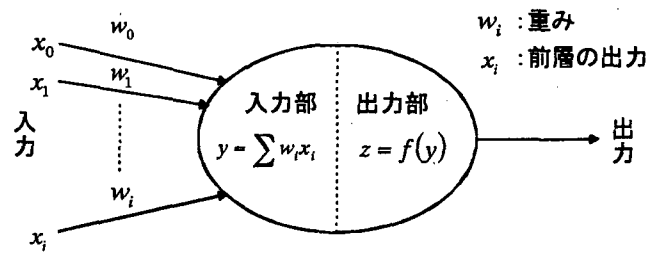


図-4 ニューロンのモデル化

るとも言える。

3. ニューラルネットワーク⁹⁾

(1) 情報伝達

脳は神経細胞を基本単位として, 細胞同士をシナプスが繋いで情報伝達を行っている。これを図-4 に示すような工学的モデル (ニューロン素子) としたものが NN である。

ニューロンは複数の入力端子と一つの出力端子を有しており, 内部には入力信号の積和を計算する入力関数部と入力の積和に従って出力を決定する出力関数 (しきい値関数) 部を有する。

ここで, 1つのニューロン素子が n 個のニューロンから入力を受けるとき, その出力を x_i とし, シナプス結合の強さを w_i とする。このとき i 番目のニューロンからの影響を単純に加算したものを入力とすると式(2)となる。

$$y = \sum w_i x_i \quad (2)$$

また, y が閾値 θ を超過したときに式(3)に示すインパルス u が発生 (興奮) する。

$$u = \sum_{i=1}^n w_i x_i - \theta \quad (3)$$

この時の出力関数 z は式(4)となる。

$$z = f(u) \quad (4)$$

ここで θ の代わりに w_0 を与えることで抑制入力とする場合もあり, この時のインパルスは式(5)となる。

$$u = \sum_{i=1}^n w_i x_i + w_0 (-1) = \sum_{i=1}^n w_i x_i - w_0 \quad (5)$$

(2) ネットワークの種類

NN は階層型ネットワーク (図-5) と相互結合型ネットワーク (図-6) に分類される。

階層型ネットワークはニューロン素子を階層構造で結合したものであり, 入力層に入力された信号が次々に

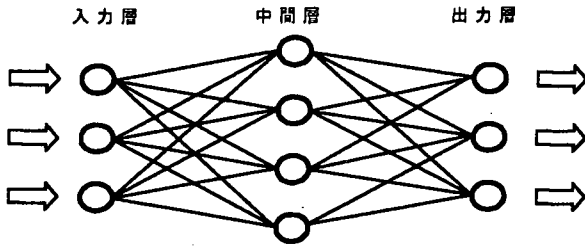


図-5 階層型ネットワーク

層を伝播しながら一方向で出力層に到達するネットワークである。一方、相互結合型ネットワークはニューロン間が相互に結合し合っており、安定した平衡状態になるまで学習するネットワークである。

本研究では階層型ネットワークを採用し、入・出力層の関係を他対1として階層型ネットワークを用いて抽出されたルールから擬似的な相互結合型ネットワーク形成についても検討している。

(3) 学習法¹⁰⁾

ネットワークの出力値と教師信号との誤差が最小になるようにノード間の結合重みを調節するもので、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バックプロパゲーション（逆伝播）と呼ばれており、階層型NNにおける最も代表的な学習法である。

ここで、各ニューロンは式(6)に示す出力（シグモイド関数）を有しており、各ニューロンに出力値は式(7)となる。

$$f(x) = \frac{1}{1 + \exp(-x)} \quad (6)$$

$$O_i^k = f\left(\sum_j W_{ij}^k O_j^{k-1} - \theta_i^k\right) \quad (7)$$

O_j^{k-1} : k-1層の各ニューロン出力, W_{ij}^k : 重み, θ_i^k : しきい値

次に時刻 t における入力信号 p の絶対誤差を式(8)とする。

$$E_p(t) = \frac{1}{2} \sum_i (T_{pi} - O_i^n)^2 \quad (8)$$

T_{pi} : 入力信号 p に対する出力ニューロン i の教師信号
ここで、ネットワークの平均誤差は式(9)となる。

$$E(T) = \frac{1}{P} \sum_p E_p(t) \quad (9)$$

BP法は式(10)に示す、誤差が可能な限り小さくなるように重み w を調整する方法である。

$$\Delta W_{ij}^k(t) = -\varepsilon \frac{\partial E(t)}{\partial W_{ij}^k(t)} + \alpha \Delta W_{ij}^k(t-1) \quad (10)$$

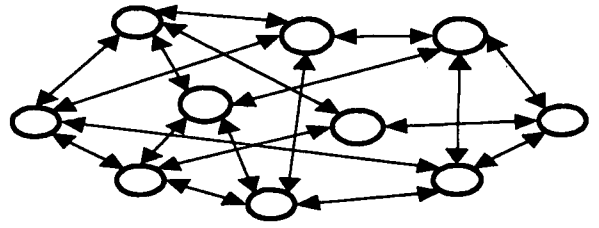


図-6 相互結合型ネットワーク

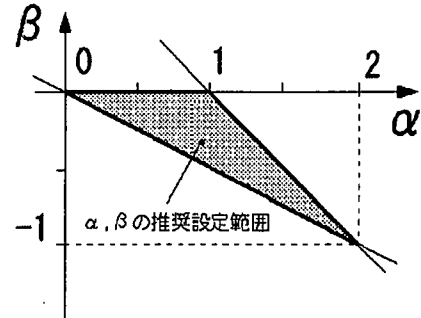


図-7 α, β の推奨設定範囲

ここで、 α は誤差の振動を減衰させ解の安定効果はあるが、学習速度の低下や局所解滞留に対する注意が必要である。

仮想インピーダンス法はBP法を拡張した学習方式であり、より高速に学習が行われるように重みの変化率を2次の振動系として扱う方法である。

この方法は式(11)に示すようにパラメータ β を含む項を追加して各ニューロンの重みを調節する。

$$\Delta W_{ij}^k(t) = -\varepsilon \frac{\partial E(t)}{\partial W_{ij}^k(t)} + \alpha \Delta W_{ij}^k(t-1) + \beta \Delta W_{ij}^k(t-2) \quad (11)$$

ここで、パラメータ β はネットワークの学習中に重みに強制振動を与え、局所解（ローカルミニマム）から脱出させる効果を有しており、 α と β は図-7の範囲内で設定することが望ましい。

次に成長側抑制学習は、データ間の法則性を抽出し易くしたものであり、仮想インピーダンス法を拡張したものである。BP法や仮想インピーダンス法では複雑な結合になってしまうため、ニューロンの結合を抽出することが難しいが、成長側抑制学習は重みの成長に抑制をかけて重要な結合だけが残るようにするため、ニューロン間の結合が明確になり、学習結果が導きやすくなる。

$$\Delta W_{ij}^k(t) = -\varepsilon \frac{\partial E(t)}{\partial W_{ij}^k(t)} + \alpha \Delta W_{ij}^k(t-1) + \beta \Delta W_{ij}^k(t-2) + S \quad (12)$$

具体的には重み更新の式(11)に新たに成長側抑制項を追加したものである。

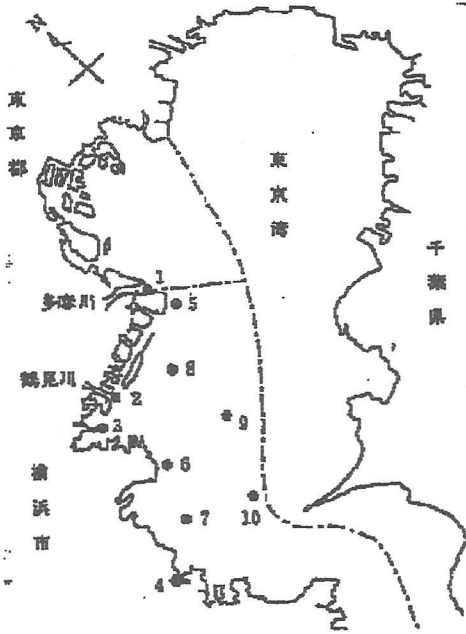


図-8 観測点位置

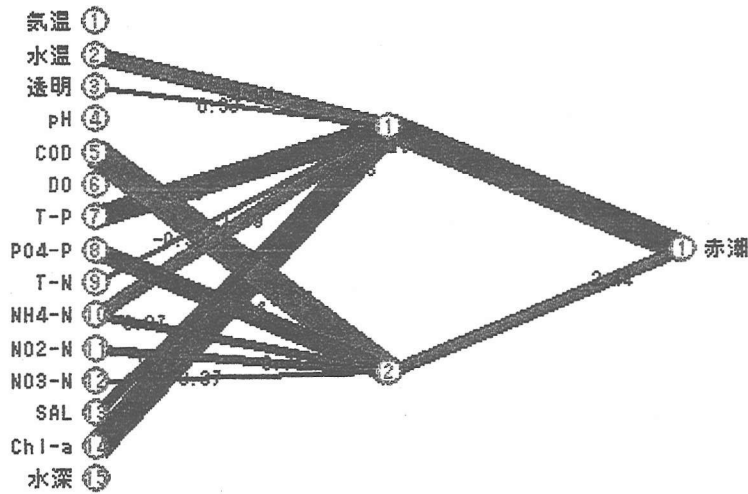


図-9 成長側抑制学習による学習結果

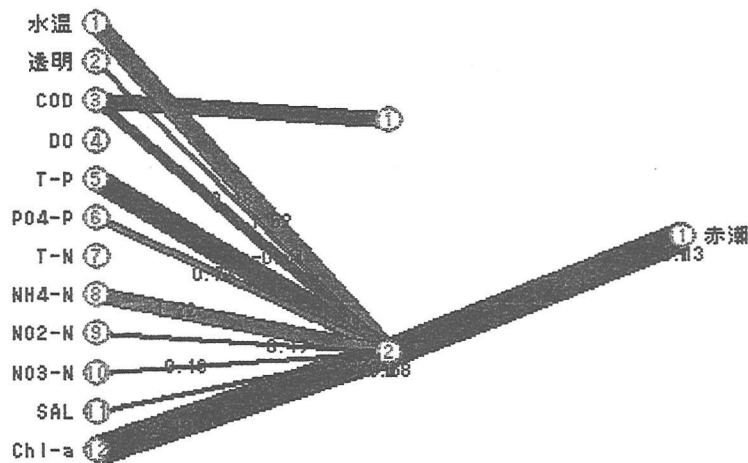


図-10 再構築後の学習結果

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	PO ₄ -P	リン酸態リン(mg/l)
I	T-N	全窒素(mg/l)
J	NH ₄ -N	アンモニア態窒素(mg/l)
K	NO ₂ -N	亜硝酸態窒素(mg/l)
L	NO ₃ -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	水深	—
P	赤潮	赤潮発生の有無

S: 成長側抑制項, s : 成長側抑制係数 (小さな値: 0.01),
 m : $k-1$ 層のノード数, W : 重み,
 θ : 閾値
 $\text{sgn}(x)$: $x < 0$ のとき -1, $x = 0$ のとき 0, $x > 0$ のとき +1 をとる関数

5. 東京湾の赤潮に関する水質観測データ

DM は表-1 に示す東京湾の 10 観測ポイント (図-8)

で得られた 16 項目の水質観測データを用いており, A~O は数値属性データ, 項目 p は赤潮発生の有無 (0 or 1 ブール属性) である。

データは 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 (=10 箇所×5 年×6 ヶ月間) である。なお, 全 300 データ中 47 データ (15.7%) で赤潮の発生が認められている。

ここで相関解析の簡潔化のため数値属性である A~O を平均値以下・以上のブール属性化し, ノイズデータは学習の収束率を向上させるためにあらかじめ削除している。

この観測項目を入力, 赤潮発生の有無を出力項目として各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 NN より求め, これより各項目と赤潮発生との相関性を「If ~ then ...」形式のルールとして抽出する。

(1) 学習 (逆解析) 結果

中間層 1 層 (ノード数 2) の階層型ネットワークにより, 気温~水深の 15 項目を入力値, 赤潮発生の有無を出力値に設定して相関解析を行った。

まず BP 法で学習を行ったが, ニューロン間の結合が複雑すぎて判別することができない。そこで図-9 に示す成長抑制学習により, 各項目 (事象) の結合関係が明確になったが, これだけでは単純なルールは抽出できない。

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

$$S = -s \frac{1}{m-1+1} \text{sgn}(W_{ij}^k(t)) \left\{ \sum_{l=1, l \neq j}^m |W_{il}^k(t)| + |\theta_i^k| \right\} \quad (13)$$

表-2 全データ(再構築後)

正の相関		負の相関	
強	弱	強	弱
Chl-a	COD	水温	PO ₄ -P
T-P	SAL	NH ₄ -N	NO ₂ -N
			NO ₃ -N
			透明

-10に示す項目「気温」、「pH」、「水深」が削除されて結合が簡素化されていることが分かる。

ここで図中,入力層~中間層~出力層が「+結合」であれば,入力と出力間には正の相関性が存在している.同様に「-結合」していれば,入力と出力間には負の相関性を示しており,相関の強さは結合の太さで表される.

図-10より「水温」、「NH₄-N」、「PO₄-P」、「NO₂-N」、「NO₃-N」、「透明」はそれぞれ平均値未満の場合に,また「Chl-a」、「T-P」、「COD」、「SAL」は平均値以上の場合に赤潮発生と相関性を示すというルールが得られる.

以上の結果から,NNを用いたDMによって得られた赤潮発生に関するルールを表-2に示す.

(2) 解析結果の妥当性の検討

それぞれの観測項目が平均値以上 (x=1) の時あるいは平均値以下 (x=0) の時にどの程度の割合で赤潮が発生したのか (条件付確率) を算出したものを表-3に示す.ここでP (赤潮有 | x=1) ,P (赤潮有 | x=0) 間に5%以上の差がある観測項目を強調して示す.またx=1のときに大きな確率で赤潮が発生しているということは,観測項目の間に正の相関性を有していること,逆にx=0のときに発生確率が大きい場合は負の相関性を示している.

ここでDMから抽出されたルールと条件付確率から得られるルールを表-4に示す.

東京湾の赤潮に関する調査報告書¹⁰⁾より赤潮発生時にCOD, Chl-aが大きいのは動植物プランクトンのデトリタスに由来し,加えてT-P, T-Nも大きい値を示すとされており,DM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おり,「水温:小」や「SAL:大」の相関も得られており,新たな知識獲得を示唆するものと考えられる.以上より,NNを用いたDMの有用性が確認された.

7. 宍道湖・中海の水質観測データ

観測データは表-6に示す宍道湖・中海の12観測ポイント(図-11)で観測された15項目の水質観測データを用いている.項目A~Oは数値属性のデータ,観測期間は1990~98年の9年間に毎月1回観測されたもので,各ポイ

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実施した結果を図

表-3 条件付確率により抽出したルール

x	x=1	x=0	A	P(A x=1)	P(A x=0)
気温	152	148	24	15.8%	16.2%
水温	160	140	22	13.8%	15.7%
透明	149	151	9	6.0%	6.0%
pH	173	127	40	23.1%	31.5%
COD	130	170	44	33.8%	25.9%
DO	132	168	35	26.5%	20.8%
T-P	121	179	32	26.4%	17.9%
PO ₄ -P	112	188	10	8.9%	5.3%
T-N	100	200	21	21.0%	10.5%
NH ₄ -N	99	201	6	6.1%	3.0%
NO ₂ -N	121	179	19	15.7%	10.6%
NO ₃ -N	90	210	6	6.7%	2.9%
SAL	200	100	33	16.5%	33.0%
Chl-a	74	226	43	58.1%	19.0%
水深	150	150	22	14.7%	14.7%
赤潮	47	253	47	100.0%	-

A → 赤潮発生
 x=0 → xの値が平均値以上
 x=1 → xの値が平均値以下

相関性が強い項目

表-4 DMと条件付確率との比較

正の相関		負の相関	
データマイニング	条件付確率	データマイニング	条件付確率
		水温	SAL
		NH ₄ -N	pH
		PO ₄ -P	
SAL	T-N	NO ₂ -N	
	DO	NO ₃ -N	
	NO ₂ -N	透明	

矛盾したルール

片方だけ手法から抽出されたルール

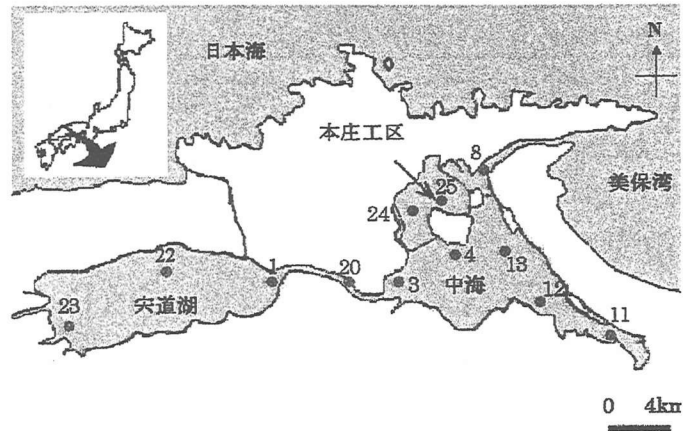


図-11 観測ポイント

ントで水面付近・湖底付近・その中間の3種類の深さで観測を行っているため,データ総数は3,888 (=12箇所×3(深さ)×9年×12ヶ月間)となる.

データは同様に,A~Oを平均値以下・以上のブール属性化を行う.また今回は出力項目(赤潮発生の有無)が明確に設定できないため,全観測値が入力であり出力項でもある.

したがって各データ間の相関(因果)を探索するためには相互結合型NNが有効と考えられるが,相互結合型

表-5 観測項目

		説明
A	trans.	透明度(m)
B	w-temp.	水温(°C)
C	pH	-
D	sal.	塩分濃度(%)
E	DO	溶存酸素量(ppm)
F	turb.	濁度
G	SS	懸濁物質(mg/l)
H	chl-a	クロロフィル a(ppb)
I	COD	化学的酸素要求量(ppb)
J	T-P	全リン(ppb)
K	PO ₄ -P	リン酸態リン(ppb)
L	T-N	全窒素(ppb)
M	NO ₂ -N	亜硝酸態窒素(ppb)
N	NO ₃ -N	硝酸態窒素(ppb)
O	NH ₄ -N	アンモニア態窒素(ppb)

は学習法が煩雑である。そこで各項目を1つずつ出力項目、残りの14項目を入力項目に設定して相関解析を行う。

ここで階層型ネットワークによって抽出した単純な「If ~ then ...」を組み合わせて、擬似的な相互結合型ネットワークを

表-6 相関性抽出表

全体 17以上を抽出

	trans.	w-temp.	pH	sal.	DO	turb.	SS	chl-a	COD	T-P	PO ₄ -P	T-N	NO ₃ -N	NO ₂ -N	NH ₄ -N
trans.	-	4	5	6	3	9	3	2	2	3	3	4	3	3	3
w-temp.	2	-	5	2	11	3	1	2	3	4	4	2	5	5	2
pH	4	5	-	8	6	5	5	5	3	2	3	5	7	5	6
sal.	2	0	5	-	2	3	6	2	4	5	2	7	8	5	2
DO	3	11	5	2	-	3	3	5	5	3	7	2	8	1	1
turb.	9	4	5	2	2	-	8	7	5	8	4	2	1	2	1
SS	7	6	3	2	3	10	-	7	4	2	1	5	3	4	3
chl-a	5	4	4	5	3	6	7	-	9	4	4	5	7	3	6
COD	1	3	3	4	3	5	3	10	-	8	4	2	6	6	6
T-P	4	3	0	4	4	7	2	2	4	-	12	3	7	2	2
PO ₄ -P	1	2	3	2	10	1	1	3	4	12	-	1	3	0	5
T-N	6	2	4	7	1	3	1	3	0	6	1	-	6	1	9
NO ₃ -N	2	7	5	3	3	2	0	4	0	3	2	4	-	11	2
NO ₂ -N	4	4	1	6	3	3	3	3	4	0	3	3	11	-	4
NH ₄ -N	3	2	5	1	5	0	3	7	3	2	3	9	4	4	-

作成して項目間の複雑な因果関係の解析を試みている。さらに観測ポイント別・季節別・水深別の解析、場所や季節・水深による相違点の解析を試みている。

(1) 相関解析 (逆解析) の結果

中間層1層 (ノード数2) を有する階層型ネットワークで、表-5に示したtrans.~NH₄-Nの各観測項目を1つずつ出力、残りの14項目すべてを入力にして成長抑制学習を行い、相互結合型ネットワークを構築する。

全観測ポイントにおいて学習を行い、以下の条件を満たすルールを表-6に示す。

- a) 全体 (12points×2 (対角の項目))
 - 24 個中 17 個以上 (71%)
- b) 水深 (Layer1, Layer6 各 6 points×2)
 - 12 個中 9 個以上 (75%)
- c) 場所 (宍道湖 4 points×2, 中海 8 points×2)
 - 8 個中 6 個以上 (75%), 16 個中 12 個以上 (75%)
- d) 季節 (夏 (6~8月), 冬 (12~2月) 各 6 points×2)
 - 12 個中 9 個以上 (75%)

また全観測ポイントの解析から得られた結果と水深・場所・季節の解析結果を表-7に示す。

表-7 全体・深さ別・場所別の解析結果

全体(全ポイント)			
rans.-turb.	chl-a-COD	w-temp.-DO	T-P-PO ₄ -P
DO-PO ₄ -P	T-N-NH ₄ -N	Turb.-SS	NO ₃ -N-NO ₂ -N
深さ別			
水面付近(Layer1)		湖底付近(Layer6)	
trans.-turb	COD-T-P	w-temp-DO	chl-a-COD
w-temp.-DO	T-P-PO ₄ -P	ph-sal	T-P-PO ₄ -P
turb.-SS	NO ₃ -N-NO ₂ -N	DO-PO ₄ -P	T-N-NH ₄ -N
		Turb.-SS	NO ₃ -N-NO ₂ -N
場所別			
宍道湖(1,2,3)		中海(3,8,11,24)	
rans.-turb.	T-P-PO ₄ -P	w-temp.-DO	chl-a-COD
w-temp.-DO	T-N-NO ₃ -N	sal-T-N	T-P-PO ₄ -P
w-temp.-NO ₃ -N	T-N-NH ₄ -N	DO-PO ₄ -P	NO ₃ -N-NO ₂ -N
sal.-NO ₃ -N	NO ₃ -N-NO ₂ -N		
turb.-SS			

新たに現れた相関性

表-8 季節別の解析結果

夏(6,7,8月)	冬(12,1,2月)
w-temp-SS	trans-turb
sal-COD	trans-NO ₂ -N
DO-NH ₄ -N	w-temp-DO
turb-T-P	ph-T-P
turb-NH ₄ -N	sal-NH ₄ -N
SS-chl-a	chl-a-COD
chl-a-PO ₄ -P	NO ₃ -N-NO ₂ -N
T-P-PO ₄ -P	
NO ₃ -N-NO ₂ -N	
NO ₂ -N-NH ₄ -N	

共通に表れた相関性

表-7より、水深別・場所別解析では全体から得られた観測データ間の相関関係とほぼ同様な結果が得られおり、水深別解析では水面付近 (Layer1) 「COD-T-P」、湖底付近 (Layer6) 「pH-sal」が相関を示した。

さらに場所において宍道湖で「w-temp-NO₃-N」、中海で「sal-T-N」に対して相関が抽出された。ここで季節別の解析結果を表-8に示す。夏冬のデータではデータ数が少なかったことより、共通なルールは「NO₃-N-NO₂-N」以外現れなかった。

全ポイントで相関が強いルールは、「trans-turb」、「w-temp-DO」、「DO-PO₄-P」、「turb-SS」、「chl-a-COD」、「T-P-PO₄-P」、「T-N-NH₄-N」、「NO₃-N-NO₂-N」の8項目であり、これらは一般的な水質観測項目間の相関関係とほぼ一致している。ここで抽出されたルールは常識的な相関性であるが、このような明確な出力項目 (結果) がないデータに対する解析法としてNNを用いたDMの有用性を確認した。

(2) 擬似相互結合型ネットワーク

全体解析で強い相関を示した「trans-turb」、「turb-SS」と「w-temp-DO」、「DO-PO₄-P」、「T-P-PO₄-P」によ

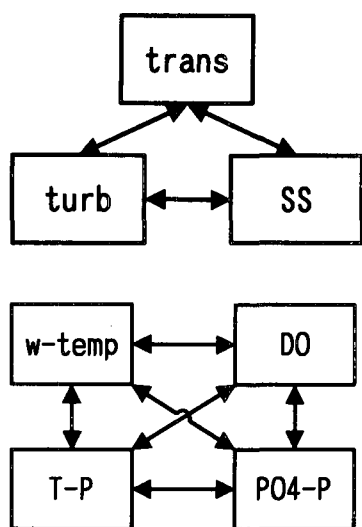


図-12 擬似相互結合型ネットワーク

り,擬似的な相互結合型ネットワークを構築すると図-12となる。ここで「trans - SS」,「w-temp - T-P」,「DO - T-P」,「w-temp - PO₄-P」は再学習の結果において相関を示すものとされた項目で,観測項目間における相関の正・負の関係も一致している。

8. 結論

本研究はデータマイニングを広い意味で逆問題として捉えて,実データを用いたニューラルネットワークによる観察事象間の集成的な相関解析を試み,以下に示す結論が得られた。

- (1) 各観測項目のブール属性化,ニューラルネットワークの成長抑制学習により,観測事象間のルール抽出が容易になった。
- (2) 赤潮発生に対して「Chl-a」,「T-P」は強い正の相関「COD」,「SAL」は弱い正の相関,また「水温」,「NH₄-N」は強い負の相関,「PO₄-P」,「NO₂-N」,「NO₃-N」,「透明」は弱い負の相関を有している。
- (3) 宍道湖・中海に対して「trans-turb」,「w-temp-DO」,「DO-PO₄-P」,「turb-SS」,「chl-a-COD」,「T-P-PO₄-P」,「T-N-NH₄-N」,「NO₃-N-NO₂-N」の8項目に強い相関性があり,その中の「trans-turb-SS」,「w-temp-DO-PO₄-P-T-P」から相互結合型

ネットワークを構築することができた。

また,今後の課題として数値データの簡便な前処理,時間的な解析手法に対する検討が挙げられる。

参考文献

- 1) Pieter Adrians and 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) 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.
- 3) 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
- 4) 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
- 5) Fayyad U.M., Piatesky-Shapiro G., Smyth P., Uthurusamy R.: Advances in knowledge Discovery and Data Mining, MIT Press, 1996.
- 6) 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.
- 7) 須藤敦史, 星谷勝: 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, 応用力学論文集, Vol.2, pp.83-90, 1999.
- 8) 久保司郎: 逆問題 (計算力学とCAEシリーズ), 培風館, 1992.
- 9) (社) 日本工業技術振興協会: (ニューロコンピューター研究部会), ニューロコンピューティングの基礎理論, 海文堂出版, 1990.
- 10) NEUROSIMTM/Light, 富士通, 1996
- 11) 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.
- 12) Edward A. Laws 著, 神田穰太・神田玲子訳: 水環境の基礎科学, 技法堂出版, 1996.

(2001年4月20日 受付)