CS-108　ニューラルネットワークを用いたデータマイニングの検討

㈱地崎工業　正会員　須藤　敦史
武蔵工業大学　正会員　星谷　勝
日本ミクニア　正会員　市村　康

1．はじめに

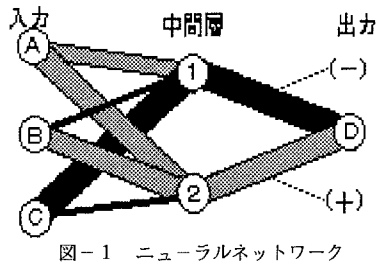
データベースを有効に活用する解析技術の必要性が高まっており，膨大なデータの分析を行う汎用ツールの開発が進められている．特に多様な種類のデータからその相関ルールの発見を目的としたデータマイニングに関連する研究が各分野で盛んに行われている．データマイニングとは従来の手法では見いだすことのできなかったデータ間の関連性や規則を導くものであり，決定木・ニューラルネットワーク・遺伝的アルゴリズムなどの学習ツールを用いた種々の手法が提案されている．そこで本研究ではニューラルネットワークを用いたデータマイニングに着目し，東京湾における水質調査データと赤潮発生の相関性の抽出を試みている．

2．データマイニング

データマイニングとは「知識の発掘」を意味し，膨大なデータベースから価値ある情報を引き出すことを目的として，データの選択・前処理・発掘・評価など複数のプロセスから構成されている．つまり‘相関ルールを発見する要素技術（学習ツール）’というよりも‘データ処理に関する基本的な考え方もしくはトータルシステム’であり，実用性の高いデータマイニングを行うためには個々のプロセスの効率化が必要となる．ここで決定木・ニューラルネットワーク・遺伝的アルゴリズムなどはデータの発掘（相関関係の解析）の際に用いられる要素技術である．

3．ニューラルネットワーク[1)]

脳の神経細胞は非常に多くのニューロンがシナプスにより結合している．情報はシナプスを通じて伝達され，各ニューロンの電位が変化し電位がある値<しきい値>を越えるとインパルスを発生させ情報を後のニューロンに伝達する．これを数理モデル化してネットワークを構成したものがニューラルネットワークである．またニューラルネットワークの学習は，人間の経験による行動選択と同様に教師データを用いた繰り返し学習が基本となる．



図－１　ニューラルネットワーク

本研究で用いたニューラルネットワークは図－１に示すような１つの中間層を有する３層の階層型ネットワークを用いており，水質調査データは平均値以上・以下のようなブール属性値（０or１）を実際のデータから作成している．ここでブール属性値のネットワークの結合は(＋)と(－)の線により構成され，結合の強さ（データ間の相関）は線の太さで表される．図－１におけるＡ～Ｄの結合と(＋)線と(－)線の結合関係は以下のようになる．

Ａ～①　(＋)　Ａ＝１の時①＝１，Ａ＝０の時①＝０
①～Ｄ　(－)　①＝１の時Ｄ＝０，①＝０の時Ｄ＝１
Ａ～①～Ｄ　(＋)(－)　∴Ａ＝１の時Ｄ＝０ or Ａ＝０の時Ｄ＝１
Ａ～②～Ｄ　(＋)(＋)　∴Ａ＝１の時Ｄ＝０ or Ａ＝０の時Ｄ＝１

したがって，線の太さと(＋)，(－)線を組み合わせることでＡとＤ間の０と１との対応関係を導き出すことができ，図－１では「Ａ＝１(０)の時Ｄ＝１(０)になる確率もしくはＡ＝１(０)の時Ｄ＝０(１)になる確率はほぼ同じである．」という結果が得られる．

表－１　赤潮データの説明

項目	
気温	14×300
水温	
透明	
pH	
COD	
DO	
T-P	
PO4-P	
T-N	
NH4-N	
NO2-N	
NO3-N	
SAL	
Chl-a	

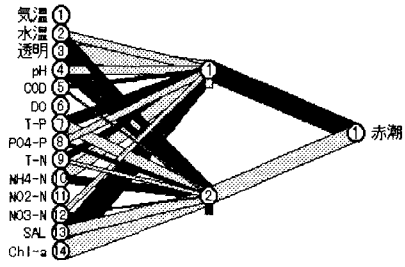
(300=10point ×5year ×6month)

４．数値解析

ニューラルネットワークを用いたデータマイニングの実問題に対する適用性を検討するために，東京湾における赤潮発生と水質調査データとの相関関係の導出に適用する．ここで用いた水質調査データの項目とデータ総数を表－１に示す．

a)基本ネットワーク

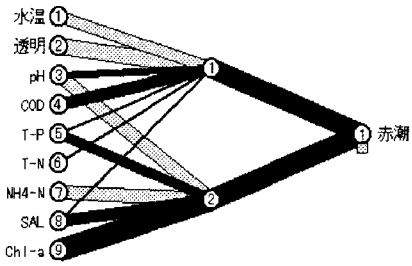
図－２に示す基本ネットワークを解析に用い，入力は１４全ての項目，出力は赤潮発生の有無のみとしている．また中間層はニューロンの少ないネットワークにおける学習精度が高いため２つとした．ここで水質調査データは各平均値以上を‘1’，平均値以下を‘0’（ブール属性値）に変換し，また赤潮の発生を‘1’未発生を‘0’としている．



図－２　基本ネットワーク（収束率 92%）

b）成長抑制学習[2)]

ニューラルネットワークの代表的な学習方法として，バックプロパゲーション法が挙げられるが，ネットワークの結合が煩雑となり項目間の相関が不明瞭になるため，本研究では成長側抑制による学習を行っている．これは重要な結合が残るように重みを調整するため，各ニューロン間の結合がより明確になる学習方法である．



図－３　再構築後のネットワーク（収束率 92%）

c）ネットワーク再構築

成長抑制学習の後にネットワーク結合の微少な項目（気温，DO，PO4-P，NO2-N，NO3-N）の削除を行って新たなネットワークを構築し，再び学習を行った結果を図－３に示す．この操作により赤潮に対する項目間の相関関係がより明確になる．またニューラルネットワークにおける学習基準として収束率が挙げられるが，本研究では項目と赤潮との相関関係を導出とその明確化を目的としているため，結合の強い項目を優先した解析が有効であると考えた．

d）解析結果

図－３より，赤潮発生と相関を有する水質調査データの項目は，「水温，透明，pH，COD，T-P，NH4-N，SAL，Chl-a」の８項目となった．

また各項目のブール属性値（０or１）と結合度の関係を表－２に示す．表－２より，各項目の赤潮発生に対する強さはChl-aが平均値以上の時と透明度が平均値以下の時，次に水温・NH4-Nが平均値以下，CODが平均値以上も何らかの関係を有すると考えられる．ここでpHは(－)(－)と(＋)(－)の結合に分かれているが，(＋)(－)の結合線が太いことより(＋)(－)の確率(平均値以下で赤潮と関連がある)が高い考えられる．

表－２　解析結果

項目	符号	データの区分	IN	赤潮発生	OUT
水温	(+)(−)	平均値以上	1	→しない	0
透明度		平均値以下	0	→する	1
NH4-N					
COD	(−)(−)	平均値以上	1	→する	1
T-P		平均値以下	0	→しない	0
SAL					
Chl-a					
pH	(−)(−)	平均値以上	1	→する	1
		平均値以下	0	→しない	0
	(+)(−)	平均値以上	1	→しない	0
		平均値以下	0	→する	1

５．まとめ

本研究は，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東京湾の水質調査データと赤潮の発生に対する相関関係の解析を行った．その結果，データに関する専門知識がなくても８つの調査項目「水温，透明，pH，COD，T-P，NH4-N，SAL，Chl-a」と赤潮との関連性は抽出することができた．これよりニューラルネットワークを用いたデータマイニングは汎用性の高い手法である．しかし，２項目以上の複合要因や各項目の時間的な前後関係については判明せず，今後の課題と言える．

［参考文献］1) 中野 馨:ﾆｭｰﾛｺﾝﾋﾟｭｰﾀの基礎,ｺﾛﾅ社,1990. 2) NUROSIM/Light ﾌﾟﾗｽ, 富士通