(VI－41)　データマイニングによる水質観測データの相関に関する考察

武蔵工業大学　学生会員　夏目誠　武蔵工業大学　正会員　星谷勝
㈱地崎工業　正会員　須藤敦史　武蔵工業大学　学生会員　佐藤大介

1. はじめに

コンピュータ技術の発達に伴い,存在するデータは毎年およそ 2 倍になっているのに対して,価値のある情報量は急速に減少しているように思われる. このようにデータ量の増加に従って, 本来ユーザーが求めている価値のある情報を探し出すことが難しくなっている. このような背景により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が可能な汎用ツールの開発が進められている. 特に多種・多用なデータから相関ルールの発見を目的としたデータマイニングに関する研究[1)]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データ解析手法では見出すことのできなかったデータ間の相関性や規則を探索するものであり, 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解析手法[2),3)]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[4)]に着目して, 宍道湖, 中海(鳥取)における水質調査データの傾向を発見する相関解析を行っている.

2. データマイニング

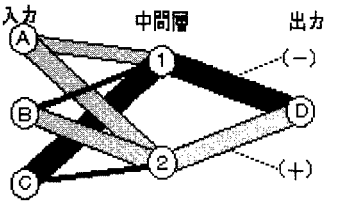
データマニイング（Data Mining）は「知識の発掘」を意味し, 膨大なデータベースから価値のある情報を引き出すことを目的として, データの選択・前処理・発掘・評価など複数の 6 つのプロセスから構成されている. つまり‘相関もしくはルールを発見する学習ツール’というよりも‘データ処理に関する基本的な考え方もしくはトータルなシステム’であり, 実用性の高いデータマイニングを行うためには個々のプロセスの効率化が必要となる.

本研究で用いているニューラルネットワークは人間の脳の結合構造を工学的にモデル化した人工生命（Artificial Life）技術であり,データの発掘（相関関係の解析）の際に用いる学習ツールである.

3. ニューラルネットワーク

本研究では, 図－1 に示すような中間層を有する 3 層の階層型ネットワークを用いており, 水質調査データは単位とディメンジョンに統一性がないため,平均値以上・以下のブール属性値(0or1)に変換している. ここで解析に用いたニューラル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. ここで図－1 における A～Dの結合と(＋)線と(－)線の結合関係は以下に示すようになる.

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



図－1　ネットワーク図

したがって,線の太さおよび(＋), (－)線を組み合わせることでAとD間の 0 と 1 との対応関係を各データ間の相関が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.

4. ネットワークの構築と学習

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, 宍道湖, 中海の水質調査データを用いた解析を行う. ここで用いた水質調査データの項目を表－1 に示す. ここで 15 個の観測項目を全てニューラル

キーワード　：　データマイニング　ニューラルネットワーク　水質データ　環境問題
連絡先　：（東京都世田谷区玉堤 1-28-1　03-3703-3111）

ネットワークの出力項目とし,解析場所は場所別(1,3,8,11,23,24), 水深別(Layer1, Layer6)に分けたため, 12ポイントであった. 中間層は入力値をブール属性値(0or1)としたのでユニット数を2個に設定し,ネットワーク総数は180個(12ポイント×15項目)となった. 学習方法は, 初めにバックプロパゲーション法を使用する. しかしバックプロパゲーション法では各データ項目間の相関関係を定量的に判別することが困難であるため,各データ項目間の結合を明確にする成長側抑制学習を用いて相関が強いものを規則(ルール)として抽出した結果, ネットワークの簡潔な構成が可能となった(図−2参照). 以下に解析手順の詳細を示す.

Step1　全観測ポイント(12ポイント)において学習し, ネットワークを構築する.

Step2　各ポイントで相関の強い項目を抽出する.

Step3　観測項目ごとに, 以下の条件にあてはまる項目を抽出する.

・全体(12ポイント)　12個中8個以上(75%)

・水深別(Layer1,Layer6 各6ポイント)　6個中4個以上(75%)

・場所別(宍道湖4ポイント, 中海8ポイント)
4個中3個以上(75%),8個中6個以上(75%)

Step4　全観測ポイント(1,3,8,11,23,24), Layer1, Layer6, 宍道湖, 中海における相関を比較した.

表—1　観測項目

trans	透明度(m)
w-temp	水温(℃)
pH	—
sal	塩分濃度(%)
DO	溶存酸素(ppm)
turb	濁度
SS	懸濁物質(mg/l)
chl-a	クロロフィルa(ppb)
COD	化学的酸素要求量(ppm)
T-P	全リン(ppb)
PO4-P	リン酸態リン(ppb)
T-N	全窒素(ppb)
NO3-N	硝酸態窒素(ppb)
NO2-N	亜硝酸態窒素(ppb)
NH4-N	アンモニア態窒素(ppb)

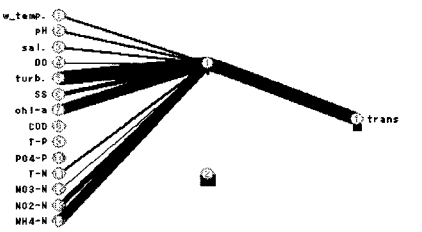


図-2　成長抑制型ネットワーク

5. 解析結果

本解析により得られた結果を表—2 に示す. 表より,水深別・位置別解析では, 全体的な観測データ間の相関関係の抽出が行えているのが分かる. また, 深さ別ではLayer1でCOD−T-P, Layer6でpH−salに対して相関を有している結果が抽出された.

加えて,場所別では宍道湖で w-temp−NO3-N, sal−NO3-N, T-N−NO3-N, 中海でsal−T-Nに対して新しい相関が抽出された. しかし, 宍道湖では2ポイントのみのデータのため, 解析個数を増加させて信用性の向上を図る必要がある.

表—2　解析結果

全体(1,3,8,11,23,24)	
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	
深さ別	場所別
Layer1	宍道湖(1,23)
Trans−turb w-temp−DO turb−SS COD−T-P T-P−PO4-P NO3-N−NO2-N	Trans−turb w-temp−DO w-temp−NO3-N sal−NO3-N turb−SS T-P−PO4-P T-N−NO3-N T-N−NH4-N NO3-N−NO2-N
Layer6	中海(3,8,11,24)
w-temp−DO Ph−sal DO−PO4-P turb−SS chl-a−COD T-P−PO4-P T-N−NH4-N NO3-N−NO2-N	w-temp−DO sal−T-N DO−PO4-P chl-a−COD T-P−PO4-P NO3-N−NO2-N

□ は新たに現れた相関性

6. 考察

ニューラルネットワークを用いたデータマイニング手法では,対象とする事象や観測データに関する専門知識が少なくても上記のようなデータ間の関連性をある程度抽出することができた. 本研究では相関の強い項目を挙げて比較したが, 相関の無い項目を挙げても新たな相関が現れると考えられる. 加えて,今後の課題として以下事項が考えられる.

1)　数値データの前処理に関する手法の確立

2)　ネットワークの中間層に関する検討

3)　入・出力を規定しない解析

4)　作業全体のシステム化・ソフト化

[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)須藤敦史,高須光郎,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,応力学論文集,Vol.2,pp.83-90,1999. 3)須藤敦史,星谷勝,市村康:ニューラルネットワークを用いたデータマイニングの基礎検討,第54回年次学術講演会,CS-108,pp.216-217,1999. 4)NEUROSIM™/L light,富士通, 1996.