CS-81

データマイニングによる非構造システムの同定に関する考察

（株）地崎工業　正会員　須藤 敦史
武蔵工業大学　学生会員　高須 光朗
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発展により大量のデータの入手が可能になってきているが，定量・定性的な量が混在する膨大なDataの解析手法が問題となっている。従来では，このようなデータに対して数量化理論などが用いられているが，相互作用が複雑なデータへの適用は難しいのが現状である．このような状況下，ある程度自動的に人間を支援し，かつ高度な情報分析手法として「データマイニング[1]」への関心が高まってきており，その有用性が期待されている．一方，土木工学においても構造物の劣化診断，交通データの因果関係追求，または環境汚染メカニズムの解明など多岐にわたり，複雑で膨大なデータを解析する手法が必要となってきている．そこで本研究は，データマイニングの概要を明らかにするとともに，条件付き確率の情報エントロピー[2]を利用した非構造システムの同定（相関関係）手法の考察を行った．

2．データマイニング

データマイニングは，「ＫＤＤ (Knowledge Discovery in Databases)」の中の役割は大きく，膨大なデータの中に存在する隠れた法則性や相関性を自動的かつ客観的に発見する手法であり，知識の発見を主眼にした意志決定のためのデータ分析手法である．ここでデータマイニングにおける分析は大きく分けて①特徴の発見，②原因の追求，③関連性の発見，④事実の観察の４つに分けられる．(Fig.1参照)

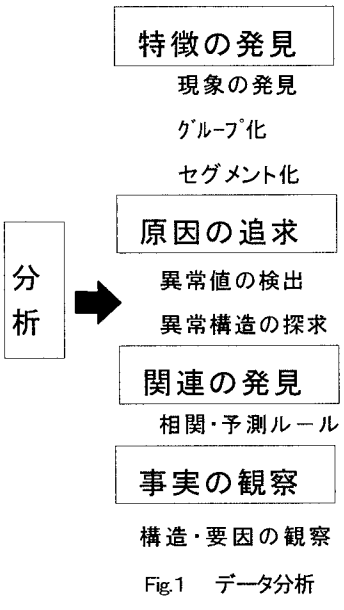

Fig.1　データ分析

またデータマイニングは，Fig.2 の示すように膨大なデータを事象の特性に対して分類（クラスタリング）し，それらのデータ間の法則性や関連性の発見をニューラルネットワーク，ＧＡなどを用いて自動的に行うものである．

ここでデータ間の法則性や関連性の探索には，属性間の直接的なネットワークとして表現するアプローチ(例えばベイジアン・ネットワーク，コウザル・ネットワーク)やエキスパートシステムの「ＩＦ～，ＴＨＥＮ …」(プロダクションルール)形式が用いられるが，従来の関係(相関)解析と大きく異なる点は，分析者がルールを与えてその法則性や関連性を解析するのではなく，データそのものから自動的に法則性や関連性を抽出するところにある．

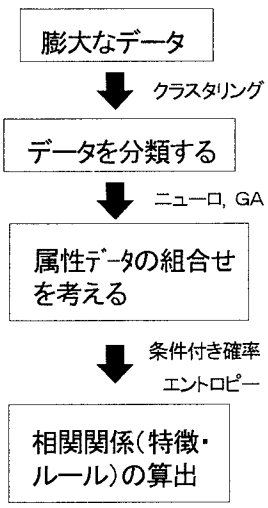

Fig.2　データマイニングの流れ

また，データマイニングにおける属性間ネットワークや「ＩＦ～，ＴＨＥＮ …」形式では，データ間の相互関係が複雑なために通常の判別関数は使いにくいため，「条件付きエントロピー」の概念を導入している．(ベイジアン・ネットワークや「IF

KEYWORD 情報エントロピー　データ処理　確率論
連絡先 住所 〒105-8488 港区西新橋2-23-1 ＴＥＬ 03-3502-2591 ＦＡＸ 03-3502-2646

～，ＴＨＥＮ …」形式が「条件付き確率」で現されることは言うまでもない．）

3．情報エントロピー

観測やアンケート調査などによりもたらされる各種のデータは，対象とする事象の未知特性に関する有用な何らかの情報をもたらし，未知特性に関するあいまい性を減少させるはずである．そこである事象の未知特性を指定したとき，データがその未知特性に関してどれだけの情報量をもたらすかが定義できるはずである．

ここで，データがもたらす情報量をC.E.Shannonの導入した相互情報量によって定義する．ここでエントロピーとは「不確定性の度合い」といった極めて抽象的なものを定量的に評価する指標であり，ある事象のエントロピーはその確率分布によって一意的に定まる．以下，具体的に相互情報量の基本的性質を検討する．

いま，離散型の確率分布を有する事象**A**，**B**を考える．

$$\mathbf{A}=\begin{pmatrix}A_1 & A_2 & \cdots & A_n\\ p_1 & p_2 & \cdots & p_n\end{pmatrix}\quad 0\le p_i\le 1\quad \sum_{i=1}^{n}p_i=1 \qquad \mathbf{B}=\begin{pmatrix}B_1 & B_2 & \cdots & B_m\\ q_1 & q_2 & \cdots & q_m\end{pmatrix}\quad 0\le q_j\le 1\quad \sum_{j=1}^{m}q_j=1$$

事象**A**のエントロピーは，式(1)のように与えられる．（**B**についても同様）

$$H(\mathbf{A})\equiv H(p_1,\cdots,p_n)=-\sum_{i=1}^{n}p_i\log p_i \tag{1}$$

ただし，$\log p$の底はeとし，$0\log 0=0$とする．$y=-p\log p$　のグラフをFig.3に示す．この関数はグラフより$p=0$と$p=1.0$で$y=0$（最小値），$p=1/e$で$y=1/e$（最大値）をとることがわかる．

以下にエントロピー$H(\mathbf{A})\equiv H(p_1,\cdots,p_n)$の性質を示す．

$$\min H(p_1,\cdots,p_n)=0 \tag{2}$$

$$H(p_1,\cdots,p_n)\ge 0 \tag{3}$$

$$\max H(p_1,\cdots,p_n)=H(1/n,1/n,\cdots,1/n)=\log n \tag{4}$$

ここで式(2)はある１つのp_iが1.0でそれ以外のp_iはすべて0（確定的現象）の場合に成立し，式(3)はエントロピーが正の値であることを示す．また式(4)はすべての事象が等しい確率で生起する場合に成立する．ここで式(4)は「不確定性の度合い」という意味からも直感的に理解できる．

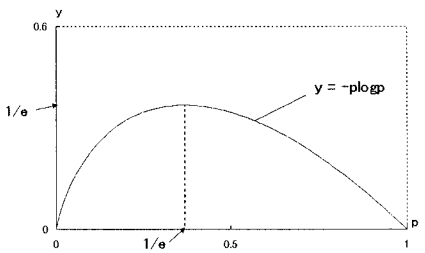

Fig.3　y＝－p log p のグラフ

次に，事象**A**と**B**の積事象**AB**を考える．この積事象のエントロピー$H(\mathbf{AB})$は式(5)のように与えられる．

$$\mathbf{AB}=\begin{pmatrix}A_iB_j & i=1,\cdots,n\\ r_{ij} & j=1,\cdots,m\end{pmatrix} \qquad H(\mathbf{AB})\equiv H(r_{11},\cdots,r_{nm})=-\sum_{i}^{n}\sum_{j}^{m}r_{ij}\log r_{ij} \tag{5}$$

ただし，r_{ij}はA_iとB_jが同時に生起する確率であり，$r_{ij}=p_iq_{j|i}$となる．$q_{j|i}$はA_iが与えられたときのB_jの条件付き確率である．また，**A**が与えられたときの**B**の条件付きエントロピー（相互情報量）$H(\mathbf{B}|\mathbf{A})$は式(6)となる．

$$H(\mathbf{B}|\mathbf{A})=-\sum_{i}^{n}p_i\sum_{j}^{m}q_{j|i}\log q_{j|i} \tag{6}$$

$H(\mathbf{AB})$，$H(\mathbf{B}|\mathbf{A})$の特性を以下に示す．

$$H(\mathbf{AB})\le H(\mathbf{A})+H(\mathbf{B}) \tag{7}$$

$$H(\mathbf{AB})=H(\mathbf{A})+H(\mathbf{B}|\mathbf{A}) \tag{8}$$

$$H(\mathbf{B}|\mathbf{A})\le H(\mathbf{B}) \tag{9}$$

ここで，式(7)と(9)において，等号は**A**と**B**が独立の場合のみ成立し，式(9)は事象**A**により**B**のエントロピーは減少する（または変化しない）ことを表している．つまり「**A**と**B**の相関が強いほど$H(\mathbf{B})$に対して$H(\mathbf{B}|\mathbf{A})$は減少する．」と解釈することができる．これより，条件付きエントロピーをデータマイニングにおいてデータ間の法則性や関連性を表す指標として用いることができる．

＜参考文献＞

1)情報・通信新語辞典97年度版，日経ＢＰ社，1996　2)有本卓：確率・情報・エントロピー，森北出版，1992