CS-84

情報エントロピーによる事例ベースからの知識発見

○武蔵工業大学大学院　学生会員　小田切　亮
武蔵工業大学工学部　正会員　皆川　勝
(株)荏原製作所　正会員　上谷　丈和

1.はじめに

エキスパートシステムの開発にあたっては，知識獲得が困難であり，またあいまいさを持った情報の取り扱いが必要である。本論文では，鋼橋疲労損傷の補修方法選定を対象問題として，データベースからの知識発見(KDD)の考え方を用いた平均圧縮情報エントロピーによるルール抽出アルゴリズムが、知識発見に有効であることを示す。

2.平均圧縮情報エントロピーによるデータベースからの知識発見

データベース中に存在する各事例を e_i とし，事例の総数を N とする。また，事例 e_i は，m 個の要素 a_{im} によって構成されている．よって事例 e_i は，次のように表すことができる。

$\{ N: e_1, e_2, \cdots, e_i, \cdots, e_N \}$　　$\{ e_i: a_{i1}, a_{i2}, \cdots, a_{il}, a_{im} \}$

データベース内から，ルールを生成する際に，条件部になり得る要素 a_{il} を y，結論部となり得る要素 a_{im} を x として，条件部と結論部になる要素を区別して表すことにする。なお，各要素は，条件部の要素にも結論部の要素にもなり得るものとする。次に，データベース内の各事例 e_i において，条件部の要素 y を持つ事例の数を n_y とし，結論部の要素 x を持つ事例の数を n_x とする。また，条件部の要素 y と結論部の要素 x を共に持つ，すなわち，ルール $y \rightarrow x$ を抽出することのでき得る事例の数を $n_{y,x}$ とする。

はじめに，結論部の要素 x の生起確率 $p(x)$ を式(1)より求め，結論部の要素が x でない場合，x の生起確率を式(2)より求める。

次に，結論部の要素 x と条件部の要素 y の結合確率 $p(x, y)$ を式(3)より求め，また，結論部の要素が x で，条件部の要素が y の結合確率 $p(x, y)$ を式(4)より求める。

また，条件部の要素 y が起こった時に結論部の要素 x が起こる，結論部の要素 x についての条件付確率 $p(x|y)$ を式(5)より求め，条件部の要素 y が起こった時に，結論部が x である条件付確率 $p(x, y)$ を式(6)より求める。

$$p(x) = \frac{n_x}{N} \quad (1) \qquad p(\bar{x}) = 1 - \frac{n_x}{N} \quad (2) \qquad p(x,y) = \frac{n_{y,x}}{N} \quad (3)$$

$$p(\bar{x},y) = 1 - \frac{n_{y,x}}{N} \quad (4) \qquad p(x|y) = \frac{n_{y,x}}{n_y} \quad (5) \qquad p(\bar{x}|y) = 1 - \frac{n_{y,x}}{n_y} \quad (6)$$

情報理論的には、通常、結論部の要素 x がデータベース内で持つ情報エントロピーを、$-log\{p(x)\}$ で示す。また、この時，結論部の要素 x を事例内の要素 $x = a_{im}$ として持つ事例が，同時に条件部の要素 y を $y = a_{il}$ として持つ場合があった時，$y \rightarrow x$ というルールが事例中に内在していることになる．この要素 y と要素 x を同時に持つ時の結論部の要素 x がデータベース内で持つ情報エントロピーは，$-log\{p(x|y)\}$ で示すことができる。$-log\{p(x)\}$ と $-log\{p(x|y)\}$ との差は、ルール $y \rightarrow x$ を生成することにより，結論部の要素 x についての圧縮された情報エントロピーと言える。この圧縮情報エントロピーのすべての事例に対する平均値を平均圧縮情報エントロピー $ACE(x, y)$ と呼び、式(7)より求める。

$$\begin{aligned} ACE(x,y) &= CE(x,y)/N \\ &= p(x,y)\log\frac{p(x|y)}{p(x)} + p(\bar{x},y)\log\frac{p(\bar{x},y)}{p(\bar{x})} \end{aligned} \quad (7)$$

キーワード：知識発見、エキスパートシステム、知識獲得、事例ベース、ネットワークシステム、エントロピー
武蔵工業大学・工学部・〒158－0087東京都世田谷区玉堤1-28-1　Tel:03-3703-3111(Ext.3252)　Fax:03-5707-2226

3．適用結果および考察

上記の平均圧縮情報エントロピーに基づく知識発見の適用例として、図-1に示す構造の因果関係で表される鋼橋疲労損傷の補修方法選定問題を選び、既存の推論システムで得られる事例群を生のデータベースとみなし、ここからルールを抽出した後、それをルールベースとする推論を実施して、抽出されたルールの有効性を検討する。平均圧縮情報エントロピーである式(3)を用いたルール抽出アルゴリズムを仮想事例より作成したデータベースに適用してルール抽出を試み、その結果 620 のルールがデータベースから抽出された。図-1には、横軸に抽出された 620 のルールをとり、縦軸に評価値である平均圧縮情報エントロピー(*ACE*値)をとった図を示す。抽出されたルールのうち、AEC 値の高いルールから順に、右に示す3通りの基準で結合係数を付与し、それぞれ新たなルールベースを構築して補修方法の選定を行った。その結果を図-3、図-4、図-5に横軸に補修方法項目の番号をとり、縦軸に各項目の可能性であるノード値をとって示す。a)の場合、図-1の様に、AEC 値による評価によって幅広く分布するルールを 0.7 と 0.4 の2段階で分類したため、ノード値による可能性の分類が十分になされていないことがわかる。これに対し、b)、c)の場合は、それぞれ 5 段階、10 段階に分類したことから、共に十分な可能性の分類を行って推論結果を提示している。

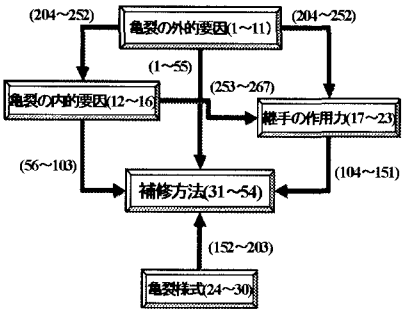

図 - 1　システム内のネットワーク構成 [1)]

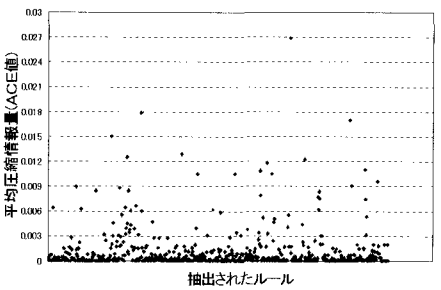

図 - 2　ACE 値により抽出されたルール

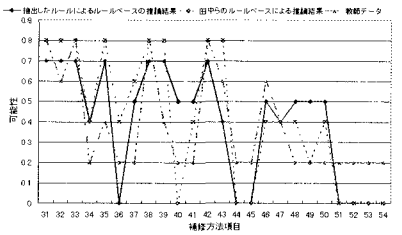

図 - 3　a)の場合の補修方法選定結果

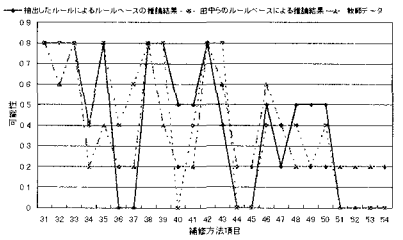

図 - 4　b)の場合の補修方法選定結果

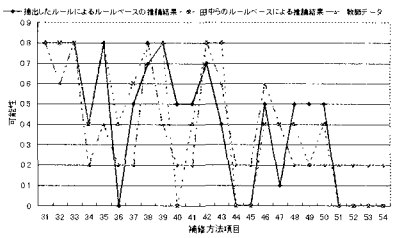

図 - 5　c)の場合の補修方法選定結果

a)全ルールを 5 分割し、上位 40%を 0.7、次の 40%を 0.4 とした場合

b)全ルールを 5 分割し、上位から 20%毎に 0.8, 0.6, 0.4, 0.2 とした場合

c)全ルールを 10 分割し、10%毎に 0.8 から 0.1 までとした場合

なお、3 通り全てにおいて、下位 20%のルールについては切り捨てた

4．おわりに

知識獲得の初期における知識獲得問題解決のため，データベースからの知識発見（KDD）の手法に平均圧縮情報エントロピー（ACE 値）を用いてルール抽出を試みた。また，抽出したルールの有用性を検討するため，抽出したルールを用いて鋼橋疲労損傷の補修方法選定のためのルールベースを構築し，実際の事例を適用して補修方法選定を試みた。この結果，推論結果は，田中らによるルールベースでの推論結果と同等の結果を示していることから，知識発見におけるルール抽出に ACE 値が十分有効であると考える。

〈参考文献〉　1)田中成典：橋梁工学への知識情報処理技術の応用に関する研究、関西大学学位論文、pp. 25-248、1996. 9.　2)Bing Leng and Bruce G Buchanan:Using Knowledge-assisted discriminant analysis to generate new comparative terms;Artificial Intelligence and Statistics IV,Springer-Verlag,pp.479-487.1993