CS-109 データマイニングによる相関ルール抽出における最小支持度の影響に関する一考察

武蔵工業大学　正会員　皆川　勝

1.はじめに　本研究では，データマイニング手法で最も一般的に用いられているアルゴリズムであるアプリオリ・アルゴリズムをファジイ属性値問題に適用してルールを抽出する方法を述べる．次に，鋼橋疲労損傷の補修方法選定に対して本方法を適用し，最小支持度および最小確信度によって抽出されるルールにどのような傾向が現れるかを考察した．

2.データベースからの相関ルールの抽出　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．

$$\{N: e_1, e_2, \cdots, e_i, \cdots, e_N\}$$

また，事例 e_i は，m 個の要素によって構成されているので，事例 e_i は次のように表すことができる．

$$\{e_i: a_{i1}, a_{i2}, \cdots, a_{im}\}$$

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を y，結論部となり得る要素を 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．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．このとき，相関ルール $y \rightarrow x$ の支持度 *support*($y \rightarrow x$)および確信度 *confidence*($y \rightarrow x$)はそれぞれ次式により定義される[1]．

$$support(y \rightarrow x): p(x,y) = n_{y,x}/N \quad (1) \qquad confidence(y \rightarrow x): p(x \mid y) = n_{y,x}/n_y \quad (2)$$

相関ルールは，第１ステップで最小支持度を満足するルール（以後，中間ルールと呼ぶ）を抽出し，第２ステップでその中から最小確信度を満足するルールを導出することで得られる．一般に第１ステップはデータベースを繰り返し検索するための負荷が第２ステップに比較して顕著に大きくなるため，種々のアルゴリズムが提案されている．

3.アプリオリ・アルゴリズム　アプリオリは現在最もひろく引用されるアルゴリズムであり，項目数（ｋ）とした場合の処理は以下のようになる．

（１）　項目数（ｋ－１）の抽出されたルールから項目数（ｋ）の候補をすべて作成する．

（２）　データベースを検索して支持度を計算する．

（３）　最小支持度を満足するもののみを抽出して，項目数（ｋ）のルールとする．

4.ファジイ属性値に対する適用　データマイニングで対象とする事象では通常，true/false という形ですべての項目の属性値が与えられるが，知識は通常あいまいさを持つことから，true/false というクリスプに表現できるとは限らない．そこで，ここでは[0,1]という属性値を持ちうる属性間の相関ルールを抽出するために，アプリオリを修正して用いることとした．上記の式（１）および（２）に示したように，通常複数の項目間の相関ルールを抽出する際には，単にその項目が true である事例をカウントするのみである．ここでは，ファジイ属性値に対応可能なように，事例 e_i の属性値 $a_{ij}\,(j=1,m)$ を用いて *support* に対する寄与の程度を示す指標として，　ルール寄与率を $\prod_k a_{ik}$（ｋは抽出する相関ルールの要素となる全項目をとる）で定義し，それの全事例に対する平均を支持度とすることとした．

5.ルール抽出事例　鋼橋疲労損傷の補修方法選定に対して本方法を適用し，最小支持度および最小確信度によって抽出されるルールにどのような傾向が現れるかを考察する．データベースから有効な知識を抽出することが可能であるかを検討するため，田中による鋼橋疲労損傷の補修方法選定システムのルールベース[2]を基に，総数2695の仮想事例を作成し，これをデータベースとした．すなわち，図-1 に示すように亀裂の内的要因，外的要因，継手の作用力，亀裂様式の 30 個の入力情報項目について，起こり得る可能性のある全組み合わせ，2695 通りを仮に発生した事実とし，これを仮想事実とした．この仮想事実に対して皆川らのルールベース洗練機能付推論システム[3]

キーワード：データマイニング，相関ルール，支持度，エキスパートシステム

連絡先：武蔵工業大学工学部土木工学科（〒158-8557 世田谷区玉堤 1-28-1, TEL/FAX:03-5707-2226）

を用いて補修方法を推定し，仮想事実と組み合わせて仮想事例とした．こ推定された補修方法の可能性は[0,1]の実数値をとる．

補修方法を直接の結論部とするルールの抽出を試みた．最小支持度が 0.07 以上の場合には必要とするルール数を抽出しないことが分かったため，最小支持度は 0.01 から 0.07 の範囲で変化させた．図-2 には，各最小支持度において抽出された中間ルール数を示す．また，図-3 には抽出された中間ルールから目標数のルールを抽出した際の最小確信度を示す．最小支持度を増加させることによって，中間ルール数は単調に減少し，最小確信度も最小支持度が 0.05 を超えると減少した．対象としたルールは Level-1 から Level-4 まで４段階の強さを設定されているが，そのそれぞれのレベルのルールおよび全ルールについて，ルールの抽出率を求めた．図-4 にその結果を示す．全体では，最小支持度を変化させても抽出率に有意な変化は見られない．しかし，Level-4 に見られるように，高い信頼度を持つルールの抽出率は最小支持度が 0.05 を超えると 100%となりきわめて効果的にルールが抽出されていることがわかる．一方，低い信頼度のルール抽出率は逆に低くなった．図-1 に示したように仮想事例を作成するために用いたルールベースでは，直接補修方法を結論部にもつルールと間に他の仮説をはさんで多段推論になっているものがある．そこで，それらを区別して，ルール抽出率を再評価した．図-5 は多段推論となるルールの条件部となる亀裂の外的要因に関連するルールの抽出率であり，図-6 はそれ以外のルールの抽出率である．図-5 の場合，最小支持度が小さいと高い信頼度を有するルールの抽出率が極端に低くなっているのに対して，図-6 の場合には，いずれの最小支持度においても Level-4 のルールの抽出率は 100%となった．多段推論ルールに対する抽出率の低下については，図-3 に示した最小確信度の低下が影響しているものと考えられるが，これについては，ルールの項目数ｋを３以上に設定して再度ルール抽出をすることで改善されると思われる．

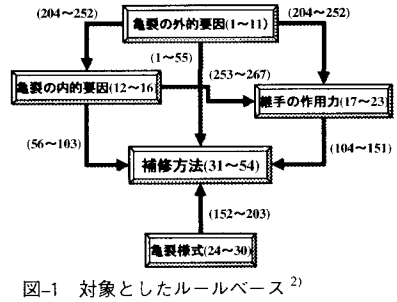

図-1　対象としたルールベース[2)]

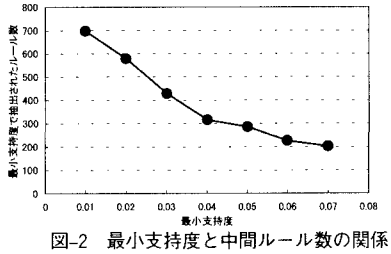

図-2　最小支持度と中間ルール数の関係

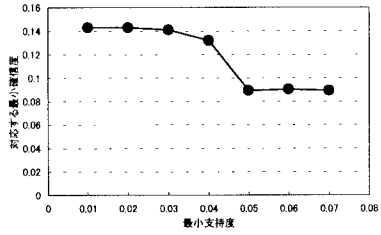

図-3　最小支持度と最小確信度の関係

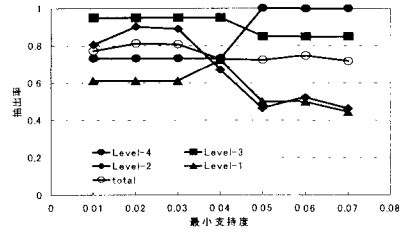

図-4　全ルールの抽出率

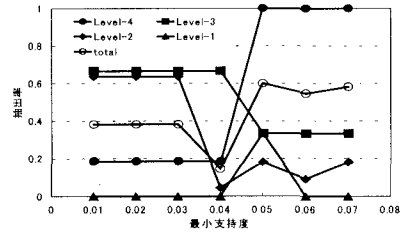

図-5　亀裂の外的要因に関連するルールの抽出率

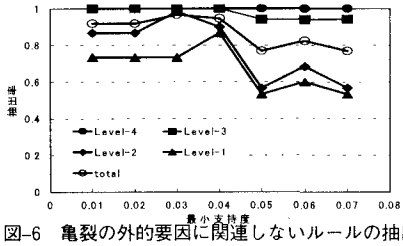

図-6　亀裂の外的要因に関連しないルールの抽出率

参考文献 1)喜連川優：データマイニングにおける相関ルール抽出技法，人工知能学会誌，Vol.12，No.4，pp.19-26，1997.7.，2)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，1996.9.，3) 皆川勝・佐藤茂・上谷丈和：事例ベースを援用した知識洗練機能付きエキスパートシステムの開発，土木学会論文集，No.595/VI-39, pp.67-76, 1998.6.