鋼橋点検データからの知識抽出のための ラフ集合論を用いたアプリケーション開発

Development of an Application Using Rough Set Theory for Knowledge Extraction from Bridge Inspection Data

江本　久雄*・絹谷　一郎**・三角　俊介***・河村　圭****

Hisao EMOTO, Itirou KINUTANI, Syunsuke MISUMI AND Kei KAWAMURA

In this study, a rough set theory application is developed in order to extract knowledge form bridge inspection data. Recently, maintenance technology for infrastructure has been notable, and also information technology such as database systems has been grown rapidly. Bridge inspection data has stored in database system. Then, it is important to evaluate from data.

As a result of developing application, it is possible to extract knowledge such as rules and reduce attributes from bridge inspection data.

Key Words: Rough Set Theory Application, Datamining, Knowledge extraction

1. はじめに

社会基盤構造物は，豊かな持続可能な社会を継続していくための重要な要素の一つである．このような構造物は，管理者により日常的に点検され，維持管理されている．また，近年の情報技術の発展により点検結果が電子データとして保管され始め，容易に大量のデータの追加・削除・検索などができるようになっている．しかしながら，膨大な点検データのために，その結果の評価や知見を発見することが困難な状態に陥っている．さらに，点検技術者の減少が考えられ，点検の効率化のために点検項目を絞り込みたいという期待もある．このように大規模なデータからの知識発見やルール抽出は，データマイニング[1]と呼ばれ，金融業や小売業[2]などで適用され，土木分野においても，その応用[3),4)]が試みられている．データマイニングの手法としては，統計的な手法や決定木など種々の方法があるが，著者らも①点検項目を識別可能な状態で減らせる，②ルールがリストされるといった点を考慮し，ラフ集合理論[5]の適用[6]を試みている．既往の研究において，ラフ集合を適用する際には，ある特定の点検データに対応した専用の実験ツールとして，アプリケーションを作成していた．そのため，汎用性もなく，非効率であった．さらに，属性値の比較には整数型を用いていたために，文字列の入力データの場合は変換を行う必要があった．

そこで，汎用的な入力形式によりラフ集合を適用できるアプリケーションの開発を試みた．本研究では，鋼橋の点検データから本アプリケーションにより知識の抽出を行い，その結果とアプリケーションの有効性について検討を行った．

*　博(工)　(有)ミツワ電器　情報システム開発室　室長（〒755-0002 山口県宇部亀浦2丁目4-1）

**　㈱リョーセンエンジニアズ 技術計算センタ 構造解析課
（〒733-0036広島市西区観音新町1丁目20番24号リョーコー・センタービル4F）

***　山口大学 工学部知能情報システム工学科　（〒755-8611 山口県宇部市常盤台2丁目16-1）

****　博(工)　山口大学 工学部知能情報システム工学科助手　（〒755-8611 山口県宇部市常盤台２丁目 16-1）

2. ラフ集合論について

2.1 ラフ集合の基本概念

ラフ集合[5),7)]は，1982年にポーランドの計算機科学者 Zdzislaw Pawlak によって提案された識別不能性のもとでの集合の記述に関する数学的理論である．その基礎概念は，類別と近似である．我々が外界の情報に対して認識を行う際，頭の中ではそれらの情報における主語（対象物）を属性に従い類別している．対象物がこの部類分けに対して同じであった場合，それらの対象物は識別できない同じ物として認識されることとなる．ラフ集合では，その識別不能関係を利用して，対象を識別するのに必要な最低限の属性の集合（縮約）や，対象が所属するクラスを識別する簡潔なルールを導き出す方法を与えている．

2.2 識別行列による縮約化

縮約の計算法[7)]は種々提案されているが，ここでは識別行列による計算法を述べる．対象に関するデータは，複数の属性とそれらの値により構成され，表として決定表$(U, C\cup D, V, \rho)$が定義される．ここで，Uは全体集合，Cは条件属性，Dは決定属性，Vは属性のとる値の集合，ρは対象と属性に対して属性値を割り当てる関数である．また，属性集合Q $(C\cup D)$と属性qが与えられたとき，$q\in Q$であることを論理式$\mathrm{El}_Q(q)$と示す．Qが任意の(i, j)に対して次式を満足するとき，

$$\bigwedge_{i,j:i>j}\ \bigvee_{q\in\delta_{ij}} El_Q(q) \tag{1}$$

Q内の全ての条件属性を用いることにより，条件属性集合Cにより，決定属性集合Dを判定することができる．さらに，この論理式を次式のように主加法形に変形する．ただし，吸収律により論理式の長さが極小な連言項のみを含んでいるものとする．

$$\begin{aligned}&\bigwedge_{i,j:i>j}\ \bigvee_{q\in\delta_{ij}} El_Q(q)=\\&\left(El_Q(q_1)\wedge\cdots\wedge El_Q(q_r)\right)\vee\left(El_Q(q_{r+1})\wedge\cdots\wedge El_Q(q_z)\right)\vee\cdots\vee\left(El_Q(q_{t+1})\wedge\cdots\wedge El_Q(q_u)\right)\end{aligned} \tag{2}$$

Qが式(2)を満たすための必要十分条件は，属性集合$\{q_1, ..., q_r\}$, $\{q_{r+1}, ..., q_s\}$, $\{q_{t+1}, ..., q_u\}$のいずれかを含めばよく，また，これらの属性集合は，吸収律により極小な連言項であるので縮約となる．

2.3 決定行列による決定ルールの抽出

決定表$(U, C\cup D, V, \rho)$が与えられたとき，決定属性集合の属性値に基づき対象の集合がp個の決定クラス$D_k(k=1, 2, ..., p)$ に分割される．このとき決定クラスD_kに応じて決定行列は，式(3)のように定義される．

$$M_{ij}^k=\left\{(a,\rho(x_i,a))\mid \rho(x_i,a)\neq\rho(x_j,a)\right\},\ i\in K_k^+,\ j\in K_k^- \tag{3}$$

ただし，$K_k^+=\left\{i\mid x_i\in C_*(D_k)\right\}, K_k^-=\left\{i\mid x_i\notin D_k\right\}$と定義する．

M_{ij}^kは，$x_i\in C_*(D_k)$と$x_j\notin D_k$のとき，x_iとx_jの値が異なる属性とそのx_iの値を示している．つまり，x_iとx_jが帰属する決定クラスが異なるとき，値が異なる属性と，その値がどのような場合にD_kと判定されているかを表している．したがって，次式

$$L(M_{ij}^{k}) = {}'\rho(x,a_1) = \rho(x_i,a_1)' \vee {}'\rho(x,a_2) = \rho(x_i,a_2)' \vee \cdots \vee {}'\rho(x,a_m) = \rho(x_i,a_m)' \tag{4}$$

が真となれば，対象 x が負事例 x_j と同じ条件属性の値をとらないことを表している．さらに，全ての $j \in K_k^-$ に対して，次式

$$\underset{j\in K_k^-}{\wedge} L(M_{ij}^{k}) \tag{5}$$

が成立すれば，いずれの負事例とも条件属性の値が同じにならないので，D_k に帰属すると判定しても矛盾しない．また，全ての正事例 $x_i \in C_*(D_k)$ に対しても同様に考慮でき，決定表全体での論理式は，次式のように表される．

$$\underset{i\in K_k^+}{\vee} \underset{j\in K_k^-}{\wedge} L(M_{ij}^{k}) \tag{6}$$

式(6)を主加法形にすると

$$\begin{aligned}\underset{i\in K_k^+}{\vee} \underset{j\in K_k^-}{\wedge} L(M_{ij}^{k}) = {} & \left({}'\rho(x,a_1) = \rho(x_1,a_1)' \wedge \cdots \wedge {}'\rho(x,a_{m_1}) = \rho(x_1,a_{m_1})'\right) \\ & \vee \cdots \vee \left({}'\rho(x,a_{m_{v-1}+1}) = \rho(x_1,a_{m_{v-1}+1})' \wedge \cdots \wedge {}'\rho(x,a_{m_v}) = \rho(x_1,a_{m_v})'\right)\end{aligned} \tag{7}$$

となり，各連言項が決定ルールの条件部となる．これを，各決定クラス D_k について行えば，極小なルールの全てが求められる．

2.4　マイニング結果の評価方法

本研究では，マイニング結果を評価する指標として，支持度・信頼度・リフト値 [2)]を用いた．以下に，これらの定義についてまとめる．

①支持度

ルールの汎用性を示す指標であり，事例全体に対するルールの条件部分と決定部分を同時に満たす事例の割合で表される．ルールが事例全体の内，どれくらいを占めているかを表す．

$$S(R,D) = \frac{card([x]_R \cap D)}{card\, U} \tag{8}$$

ここで，$S(R,D)$は信頼度を，R は同値関係を，D は決定部分を，$card$ は事例数を，$[x]_R$ は条件部分が x の同値関係を，U は全事例を表す．

②信頼度

ルールの正確さを示す指標であり，ルールの条件部分を満たす事例と，ルールの条件部分と決定部分を同時に満たす事例の割合で表される．

$$C(R,D) = \frac{card([x]_R \cap D)}{card[x]_R} \tag{9}$$

ここで，$C(R,D)$は信頼度を，R は同値関係を，D は決定部分を，$card$ は事例数を，$[x]_R$ は条件部分が x の同値関係を表す．

③リフト値

同一規則のない決定アルゴリズムの事例全体に対するルールの決定部分と合致する事例全体の割合に対する，ルールの信頼度の割合である．つまり，決定部分の支持度に対するルールの信頼度の割合である．

$$L(R,D) = card\ (R,D) / \frac{card\ D}{card\ U} \tag{10}$$

ここで，$L(R,D)$はリフト値を，Rは同値関係を，Dは決定部分を，$card$は事例数を，Uは全事例を表す．

3. アプリケーション開発について

3.1 アプリケーションの目的と概要

本アプリケーションの目的は，ラフ集合論によるデータマイニングを行うことである．ただし，特定の入力データを用いず，CSV 形式の入力形式に従うことで試験的にラフ集合論の結果が得られることとする．画面設計としては，単純明快にするために図 1 に示すような SDI とした．また，解析可能なデータ数は，サンプル数，属性の数，属性値の数により異なるが，100 程度である．さらに，特徴のないようなデータでは，解析可能なデータ数は半分程度となることもある．

3.2 入力形式と出力形式

入力形式は，汎用性を高めるために CSV 形式とした．フォーマットは，表 1 のように 1 行目に項目名を，それ以降の行にサンプルデータを入力する．また，サンプルデータは，文字・数値型どちらでも利用可能とした．

出力形式は，解析日時，時間，ファイル名，サンプル名，属性を各行に出力後，表 2(a)のように縮約結果と表 2(b)のように極小決定アルゴリズムによる決定ルールを属性値毎に出力する形式とした．また，ルールに関しては，評価指標として支持度，信頼度，リフト値を出力するように

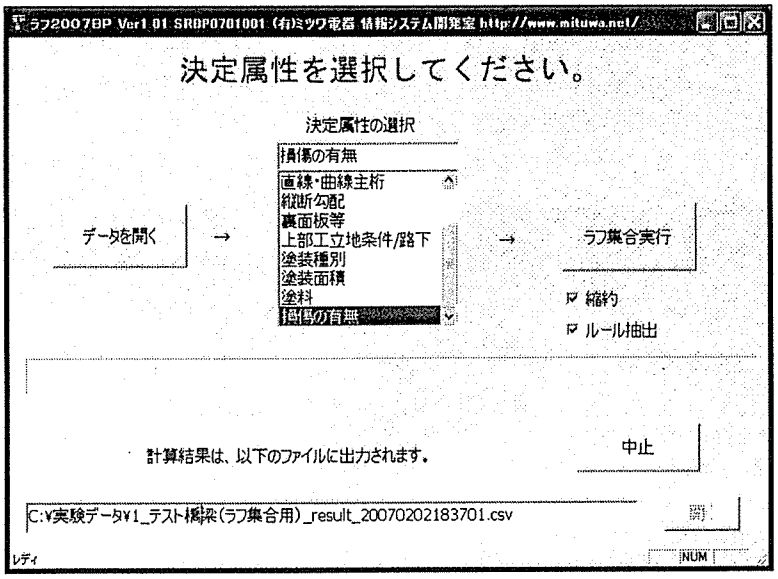

図 1　アプリケーションのメイン画面

した．ただし，決定属性値毎にルールを抽出しているために，信頼度，リフト値に関しては同値となり，評価に利用できない．ただし，信頼度は 1.0 となり，アプリケーションによって正しくルール抽出ができていることが確認できる．また，リフト値に関しては，1.0 以下の場合，条件属性に無関係で決定属性が発生していることを表しているため，リフト値が 1.0 より大きいほど，その条件属性が決定要因に必要であることが分かる．

表 1　入力データの形式

ID	条件属性名 1	条件属性名 2	～	条件属性名 5
p1	属性値	属性値	～	属性値
p2	属性値	属性値	～	属性値
⋮				
p6	属性値	属性値	～	属性値

表 2　出力データの形式

(a) 縮約結果

[縮約結果]
条件属性名 1　条件属性名 2
条件属性名 1　条件属性名 3
条件属性名 1　条件属性名 4

(b) ルール抽出結果の抜粋

[決定属性]が [属性値 A]となる ルール	条件属性 1	条件属性 2	条件属性 3	条件属性 4	全件数	条件の件数	～	結論の件数	支持度	信頼度	リフト値
属性 4 = 値 1				値 1	6	2		3	0.333333	1	2
属性 1 = 値 1 and 属性 2 = 値 1	値 1	値 1			6	1		3	0.166667	1	2
属性 2 = 値 1 and 属性 3 = 値 1		値 1	値 1		6	1		3	0.166667	1	2

4. 鋼橋点検データへの適用

4.1　点検データの概要

点検データは，条件属性数を 10，サンプル数を 84 とし，その条件属性は表 3 のように設定した．また，条件属性中の最大主桁間隔，最小主桁間隔と塗装面積は，連続値であるので表 4 に示すようにカテゴリーを分けた．さらに，決定属性としては「損傷の有無」とし，その値は，「有り」・「無し」の 2 値である．なお，本アプリケーションを検証するためにデータ量を少なくし，専門家によって架空の特徴的なサンプルデータを作成した．サンプルデータの主な特徴としては，以下のように設定した．

① 塗装種別で、塩化ゴム系の損傷はまれとした.

② 主桁では、主桁数が多いほど損傷が多いとした.

③ 塗料という項目を作り，任意に値を与えた.

表 3　点検データの抜粋

ID	主桁数	最大主桁間隔	最小主桁間隔	直線・曲線主桁	縦断勾配	裏面板等	上部工立地条件/路下	塗装種別	塗装面積	塗料	損傷の有無
p1	2	4	3	曲線主桁	0.3	化粧板(側面板あり)	河川	塩化ゴム系	3	A	無し
p2	3	3	2	直線主桁	0.3	鳩防止ネット	河川	ポリウレタン樹脂	5	B	有り
p3	2	4	3	直線主桁	0.3	鳩防止ネット	公共用地	ポリウレタン樹脂	4	B	無し
～											
p83	2	3	3	直線主桁	3.5	なし	その他用地	塩化ゴム系	3	C	無し
p84	2	2	2	曲線主桁	0.4	化粧板(側面板あり)	一般道路	塩化ゴム系	3	A	無し

*網掛け部分は，決定属性である.

表 4　最大主桁間隔，最小主桁間隔，塗装面積のカテゴリー分けの一覧

最大主桁間隔	(m)
1	0 ～ 1
2	1 ～ 5
3	5 ～ 10
4	10 ～ 15
5	15 ～

最小主桁間隔	(m)
1	0 ～ 1
2	1 ～ 5
3	5 ～ 10
4	10 ～ 15
5	15 ～

塗装面積	(m^2)
1	0 ～ 100
2	101 ～ 500
3	501 ～ 1000
4	1001 ～ 2000
5	2001 ～

4.2　解析条件とその結果

本アプリケーションで得られた縮約結果を表 5 に，「損傷の有無」が「有り」となるルールを表 6 に，「無し」となるルールを表 7 に示す.

表 5 に示す縮約結果から条件属性の数が 10 項目から 5 項目に減り，損傷の有無を判定する際には 5 項目に着目すれば識別できる．さらに，縮約結果が 2 つであるが，そのどちらともに含まれている 4 項目は「損傷の有無」を判定する際に重要な属性となる．また，縮約結果に「塗料」が残っていることから，この属性による影響もあることが分かる.

表 6 に示す「損傷有り」となるルールは，いずれも主桁数が 3 を含んでいる．これは，主桁数の属性値は 1,2,3 の 3 つであるので，主桁数が多くなることで損傷が発生しやすいことと一致する．また，「塗装種別 = ポリウレタン樹脂」は，サンプルデータの特徴とよく一致している.

表 7 に示す「損傷無し」となるルールの中で，「塗装面積 =2」はデータの中でもっとも小さい値のものであるので，「損傷無し」となる可能性が高い．また，「縦断勾配 =0.4 and 主桁数 =2」が抽出されており，「比較的勾配がついており主桁数が少ない場合，損傷も少ない」とした専門家の作成したデータの傾向を表している．さらに，意外なルールとしては，任意にデータを与えた

表 5　縮約結果

[縮約結果]
最大主桁間隔 上部工立地条件/路下 塗料 塗装面積 主桁数
最大主桁間隔 上部工立地条件/路下 塗料 塗装面積 最小主桁間隔

表 6　支持度の高い上位３つの「損傷有り」ルール

[損傷の有無]が[有り]となるルール	支持度	信頼度	リフト値
主桁数 = 3 and 塗装種別 = ポリウレタン樹脂	0.047619	1	8.4
主桁数 = 3 and 裏面板等 = 鳩防止ネット	0.035714	1	8.4
主桁数 = 3 and 塗料 = B	0.035714	1	8.4

表 7　支持度の高い上位３つの「損傷無し」ルール

[損傷の有無]が[無し]となるルール	支持度	信頼度	リフト値
塗料 = C	0.27381	1	1.135135
塗装面積 = 2	0.22619	1	1.135135
縦断勾配 = 0.4 and 主桁数 = 2	0.214286	1	1.135135

「塗料」が抽出された．これは，サンプルデータからデータマイニングによって意外な知識が発見できたと考えられる．

ここで，本研究で用いたデータは，データ量が少なく，専門家によると推測しやすい特徴的なものであるので，今後種々のデータによる検討が必要である．

5. 結論

本研究で得られた成果を以下にまとめる．

(1) 本アプリケーションを用いることでデータ量やデータの性質により制限があるものの，入力フォーマットに従えば汎用的にラフ集合の適用が可能となった．そのため，種々の点検データなどに適用が期待される．

(2) 本アプリケーションの実行結果には，支持度・信頼度・リフト値により結果の指標を示している．これにより結果の評価が行え，知識やルールを抽出したい管理者にとって有効な情報が提供できることが分かった．ただし，支持度のみ評価に有効であるので，評価指標に関して今後検討が必要である．

(3) 鋼橋の点検サンプルデータに適用した結果から，10 項目の条件属性数が半分の 5 項目に縮約され，縮約化により重要な項目が絞れた．ただし，データに特徴がないような場合は，大幅に属性数を減らすことは困難と思われ，今後の検討課題と考えている．

(4) 専門家により特徴的な点検サンプルデータを用いて本アプリケーションの妥当性について検討を行った．その結果，専門家の意見と一致する結果が得られたとともに，「損傷無し」・「損傷有り」のルールで「塗料」のような意外なルールもアプリケーションを実行することで抽出できた．

(5) 本アプリケーションは，属性値に文字列の取扱いも可能である．しかし，属性値が同じ意味でも表現が異なる場合は，別属性となる．今後，データの前処理段階の手間を大幅に改善していくためにも，テキストマイニング手法についての検討が必要である．

参考文献

1) 元田浩，津本周作，山口高平，沼尾正行：データマイニングの基礎，オーム社，2006.12.
2) 株式会社 SAS インスティチュートジャパン:データマイニングがマーケティングを変える！，PHP 研究所，2001.4.
3) 古田均，広兼道幸，田中成典，三雲是宏：橋梁の損傷要因診断事例からのラフ集合を用いたルール型知識の獲得方法，構造工学論文集，Vol.44A，pp.521-528，1998.3.
4) 古田均，木村壽夫，広兼道幸，田中成典，三雲是宏：橋梁の損傷要因診断事例からのラフ集合を用いたルール型知識の獲得および共有方法に関する研究，Vol.45A，pp533-541，1999.3.
5) 森典彦，田中英夫，井上勝雄：ラフ集合と感性，海文堂，2004. 4.
6) 加賀山泰一，河村圭，宮本文穂，田中伸也：ラフ集合の概念による橋梁伸縮継手損傷のルール型知識獲得，土木学会論文集，No.735，VI-59，pp.157-170，2003.6.
7) Z.Pawlak: Rough sets, Int. J.Inform. Comput. Sci., Vol.11, No.5, pp.341-356, 1982.