

矛盾を考慮した決定表簡約化手法による鋼橋損傷劣化ルールの抽出
Rule-type knowledge acquisition from deterioration condition of steel bridges
based on decision table abridgement considering inconsistency

脇本文	河村圭	中村秀明	宮本文穂
Fumi Wakimoto	Kei Kawamura	Hideaki Nakamura	Ayaho Miyamoto
山口大学	山口大学	山口大学	山口大学
Yamaguchi University	Yamaguchi University	Yamaguchi University	Yamaguchi University

Abstract ; As rapid progress of information technology in recent years, many organizations have been applying the database systems to complication. However accumulation extensive data made possible, accumulated data is not utilized effectively. Then, our laboratory has been trying acquisition of the rule type knowledge form accumulated data by using the decision table abridgment, which used the concept of rough sets. However, there is a problem in this technique, which cause by the handling to inconsistent data. This problem disturbs acquisition of effective knowledge. This paper proposes the solution method of the problem.

1. まえがき

近年の情報処理技術の発展に伴い、橋梁維持管理分野においても緒元や点検結果等のデータ蓄積方法は、従来の紙主体から、データベースシステムへと移行してきている[1]。大量のデータを蓄積可能になったが、データの有効な活用が十分に成されていないのが現状である。そこで本研究室では、データマイニングと呼ばれる技術を使用し、橋梁維持管理の為のデータベース内に存在するデータを解析し、橋梁の損傷劣化に関する知識の抽出を行ってきた。その際に用いたのはラフ集合の概念を用いた決定表簡約化手法と呼ばれる手法である。

本研究では、データ解析の能率を上げるため、決定表簡約化手法の解析課程において棄却されていた矛盾データを利用するように解析プロセスの改良を行った。

2. データマイニング

データマイニング[2]とは統計解析と近親性のあるデータ解析プロセスである。統計解析が適量のデータを用いて仮説の検証を行うのを主な目的としているのに対し、データマイニングでは仮説を設定することなしにデータの解析を行い、新規の仮説を発見することを主な目的としている。このため、データマイニングでは解析に大量のデータを必要とする。この統計解析との相違点である、『大量のデータ』及び『新規仮説の発見』がデータマイニングの特徴を示すキーワードである(図1)。



図1 データマイニング

3. ラフ集合

ラフ集合[3]とは、類似と近似を基本概念とし、ものごとをあらく扱うための集合論である。ラフ集合では、関心のある属性の値に従って対象と他のものを識別する。関心のある属性の値が同じであれば同じものとして扱ってよいとする。例えば「犬」と「猫」を『ほ乳類であるか否か』のみに着目してみれば、「ほ乳類」という同一のものとして扱え、また逆に『サイズ』という属性の値を用いて「犬」を細かく見ていくことで「中型犬」や「大型犬」のように詳しく類別できるようになる。このように、既知の値の差異のみでしか対象を分類できないことを識別不能性といい、これがラフ集合の最も基本的な概念となる。

4. 決定表簡約化手法

ラフ集合の概念を利用した技術として、決定表の簡約化があげられる[4]。決定表とは特定の条件下においてどのような決定を下すべきかを規定するものである。意志決定問題のほとんどが決定表による定式化が可能であるため、多くの分野において利用されている。

表1 決定表の例

決定規則	条件属性				決定属性 e
	a	b	c	d	
1	1	1	1	1	A
2	1	2	1	1	A
3	2	1	1	1	A
4	1	2	2	1	B
5	2	2	1	1	B
6	2	1	2	1	B
7	2	2	2	1	B
8	3	2	2	2	C
9	3	2	1	2	C
10	3	1	1	2	C
11	3	1	1	2	C
12	3	1	2	1	A
13	3	1	2	1	B
14	3	1	2	1	C
15	3	1	2	1	B

表2 簡約化された決定表

決定規則	条件属性			決定属性 e
	a	b	c	
1	1	×	1	A
2	1	×	1	A
3	2	1	1	A
4	1	×	2	B
5	2	2	×	B
6	2	×	2	B
7	2	×	2	B
8	3	×	×	C
9	3	×	×	C
10	3	×	×	C

決定表の一例を表1に示す。意志決定に使用する情報を条件属性と呼び、条件属性の値が決まることによって意志決定を下す決定属性の値が一意に決まる。

決定表に関わる多くの応用分野において、決定表の簡約化という問題は、最も重要な話題となってくる。決定表の簡約化の一例として、決定表内の条件属性の簡約化を挙げることができる。簡約化の結果、より少ない条件に基づいて簡約化前と同一の決定を下すことが可能となり、意志決定に不要な条件を調べる手間を省き、コストの削減をはかることが出来る。例えば表1を簡約化した表2においては、条件属性dを調べる必要が無く、また決定規則8などにみられる特定の条件では、一つの項目を調べるのみで意志決定が行えるため、その分の手間やコストを省くことが可能になる。

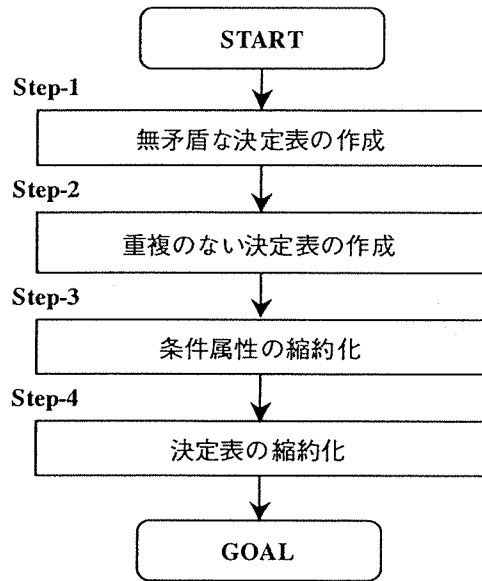


図2 決定表簡約化の流れ

決定表の簡約化手順

以下に、図2を用いて表1から表2のような簡約化された決定表を得るためのアルゴリズムを説明する。

Step-1; 無矛盾な決定表の作成

はじめに、“決定属性は条件属性に従属するか”（従属性が成り立つか否か）を決定する。この結果、矛盾のない決定規則のみが決定表に残され、以下のSTEPは矛盾のない決定規則のみについて行われる。

Step-2; 重複のない決定表の作成

次に、1つの決定表の中に複数の同一な決定規則が存在する場合は、1つの決定規則を残し、後の決定規則を決定アルゴリズムから削除する。

Step-3; 条件属性の縮約化

続いて、“条件属性の集合は不要な属性を含んでいるか否か”，すなわち、決定表の分類能力を壊すことなしに条件属性のうちのどれが削除できるか否かを決定する。

Step-4; 決定表の縮約化

最後に、決定表の最簡約形を求めるために、決定表における余計な属性値を取り除けるか否かを決定する。

この流れにより、得られた極小決定アルゴリズムは、if-thenルールとして表される。

5. 矛盾を考慮した決定表簡約化手法

決定表の簡約化が行われる際、矛盾レコードは無矛盾な決定アルゴリズムの作成 (Step-1)

の段階において、不要なレコードとして排除されることを前節で述べた。そのため Step-2 以降のルール抽出作業では無矛盾な決定表のみが使用され、矛盾を含んだレコードは全く考慮されないことになる。簡約化の手法上仕方のない処置ではあるが、ここに大きな問題が潜んでいる。ルールの抽出作業に一部のデータしか使用されないため、簡約化される前の決定表にそぐわないルールが抽出されるということである。

そこで本研究では、元のデータに忠実なルールの抽出を可能にするため、①Step-1 で削除されるデータの減少、②条件属性を簡約化の際に矛盾のある決定表も考慮に入れることを考えた。具体的な方法を以下に示す。

- ① Step-1 において矛盾する決定規則を削除する際に、発見された矛盾する決定規則が例外的なものでないか調べる。具体的には、決定表中に存在する矛盾する決定規則と同一の決定規則の数を x 、それらの決定規則に対して矛盾する残りの決定規則の数を y とし、次のことが成り立つか否かを計算する。

$$x/(x+y) \geq \alpha$$

ここで α は、決定規則が矛盾するものであるかのしきい値である。(本研究では $\alpha = 0.75$ と設定した)。これが成り立つ場合、発見された矛盾する決定規則は無矛盾であると判断される。

- ② Step-1 において、矛盾とみなされた決定規則は矛盾のない決定表とは別に残しておく。Step-4 において、条件属性の削除の際、この完全に矛盾した決定表を参照しながら削除を許可するか否かを判断する。具体的には、従来の手法で矛盾のない決定表から省略する属性値を決めた後、矛盾のない決定表、完全に矛盾した決定表を見て、両決定表中に存在する矛盾する決定規則と同一の決定規則の数を x 、それらの決定規則に対して矛盾する残りの決定規則の数を y とし、①の式に当てはめる。もし成り立つならば、矛盾は例外的であると判断し省略を認め、成り立たなかった場合省略を許可しない。

6. 実データを使った各手法の比較

既存の決定表簡約化手法（以下、旧手法）と今回提案した矛盾を考慮した決定表簡約化手法（以下、提案手法）との比較のために、阪神高速道路公団から頂いた道路高架橋の鋼桁に関する維持管理データを用いた。今回は特にさびおよび腐食の損傷進行度に関する決定表を簡約化

した場合を比較する。『橋齢』、『橋長』、『縦断勾配』、『直線・曲線主桁』、『バチ・斜形状』、『桁種別』、『主桁数』、『主桁高さ』、『最大主桁間隔』、『床版種別等』、『床版厚』、『部材』、『路下条件』の 13 項目を条件属性とし、『さびおよび腐食・損傷判定』の 1 項目を決定属性とした。そのときの入力データ数は 3103 件となる。

検証に用いる指標である信頼度、支持度について以下に説明する。

信頼度：ルールの条件属性を満たす事例に対する、ルールの条件属性と決定属性を同時に満たす事例の割合のことであり、ルールの確信度とも言い換えることができる。信頼度を定式化すると、以下の式(1)が得られる。

$$C = \frac{\text{card}([x]_R \cap D)}{\text{card}[x]_R} \quad (1)$$

ここで、 C は信頼度を、 $[x]_R$ は条件属性の一致を、 D は決定属性の一致を、 card は事例数をそれぞれ表す。

支持度：事例全体（解析用データ）に対するルールの条件属性と決定属性を同時に満たす事例の割合のことであり、ルールが事例全体の内どれくらいをカバーしているかを表す。支持度を定式化すると、以下の式(2)が得られる。

$$S = \frac{\text{card}([x]_R \cap D)}{\text{card } U} \quad (2)$$

ここで、 C は支持度を、 $[x]_R$ は条件属性の一致を、 D は決定属性の一致を、 card は事例数を、 U は全事例をそれぞれ表す。

各々の手法で得られたルール数及びルールに使われた条件属性数の 1 ルールあたりの平均値を表 3 に示す。また、旧手法と提案手法の信頼度、支持度の度数を図 3、表 4 に示す。これらは抽出されたルールを元のデータに当てはめて算出したものである。支持度は 1% を切るような低い値への偏りが甚だしく、ヒストグラムで提示しにくいいため、表の形で示した。

表 3 から、抽出されるルール数は旧手法の方が少ないことが解る。また、平均条件属性数も同様に旧手法の方が僅かながら少ない。データマイニングは、データに隠れた法則を人間に解りやすい形式で提示することを目的とするので、旧手法の方が良い抽出を行っていると言える。

表 3 ルール数・属性数の比較

	抽出ルール数	平均条件属性数
旧手法	2281 件	3.8 属性
提案手法	3320 件	4.0 属性

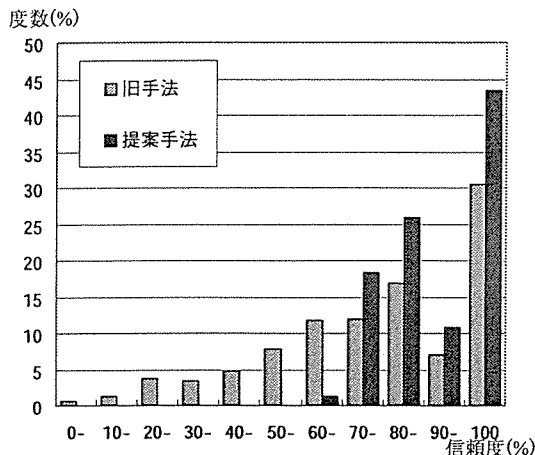


図3 信頼度による度数の分布

表4 支持度による度数の分布

支持度 (%)	旧手法 (%)	提案手法 (%)
0以上1未満	77.07	78.52
1以上2未満	12.89	13.31
2以上3未満	5.13	4.80
3以上4未満	2.32	1.71
4以上5未満	1.14	0.67
5以上6未満	0.66	0.43
6以上7未満	0.44	0.21
7以上8未満	0.22	0.21
8以上9未満	0.09	0.06
9以上10未満	0.04	0.06
平均支持度 (%)	0.7496528	0.6826436

図3から、旧手法では解析元のデータに当てはめても信頼度の著しく低いルールが抽出されるのに対し、提案手法ではコンスタントに一定以上の信頼度を持つルールが抽出できることが解る。これは劣化予測に信頼できるルールが多いことを示し良い傾向であるといえる、逆に表4によると、提案手法は旧手法よりも支持度が低くなっている。このことから提案手法のルールは旧手法より限られた条件でしか使用できないことが判明した。

次に、各々の抽出結果の中で、支持度の高かったルールの条件属性数、信頼度および支持度を上位3位まで表5にまとめた。これをみれば解るとおり、旧手法において抽出されていた信頼度、支持度の高いルールと同等のルールは提案手法でも抽出されている。このため、全体的に支持度が低くなるのは、旧手法において支持度が高いばかりで信頼度の低いルールが抽出されないからでると考えられる。

表5 支持度の高かったルール

	属性数	信頼度	支持度	
旧手法	1	4	83.91	9.41
	2	4	76.4	8.77
	3	5	83.67	8.09
提案手法	1	4	84.32	9.70
	2	4	83.91	9.41
	3	4	76.4	87.66

以上から、提案手法は多少の難点があるものの、十分に実用に耐えるものである。

7. まとめ

本研究では、ラフ集合の概念による決定表簡約化手法で問題となっていた矛盾するデータの扱いの改善を目的として、決定表簡約化手法のプロセスの変更を行った。

以下に本研究で得られた成果を示す。

- ① 既存の手法では不必要なデータとして除外されていたデータを使用し、簡約化に規制を作ることで信頼度の高いルール抽出に成功した。
- ② 信頼度を意図的に向上させるには、支持度が犠牲になり、また抽出されるルール数が多く煩雑になってしまう傾向があることが判明した。

参考文献

- [1] 加賀山泰一, 河村 圭, 宮本文穂, 田中信也: 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, pp157-170, 2003.6
- [2] Fayyad & Piatesky-Shapiro: Knowledge Discovery and Data Mining, 石川慎也訳, 1998
- [3] 中村 昭: ラフ集合—その理論と応用 第1回, ラフ集合 その概念と知識情報, 数理科学 No.373, pp78-83, 1994.7
- [4] 中村 昭: ラフ集合—その理論と応用 第2回, ラフ集合と論理・推論, 数理科学 No.374, pp86-91, 1994.8

連絡先

脇本 文
 山口大学大学院理工学科
 知能情報システム工学専攻
 〒755-8611
 山口県宇部市常盤台2丁目16-1
 TEL&FAX 0836-85-9530
 E-mail wkimoto@design.csse.yamaguchi-u.ac.jp