ラフ集合の概念を利用した決定表簡約化における属性選択手法の提案

山口大学大学院　学生会員○河上和志　　山口大学工学部　正会員　河村　圭
宇部情報システム　田中信也　　山口大学工学部　正会員　宮本文穂・中村秀明

1. はじめに

橋梁データベースには橋梁の維持管理に関する膨大な量のデータが蓄積されている．これらのデータは，大量のデータから価値ある情報を引き出す解析方法としてのデータマイニング技術により，橋梁部材の劣化・損傷に関する何らかの価値ある情報を引き出せる可能性があると考えられる．よって，橋梁維持管理の分野において，データマイニング技術は有効なものであると考えられる．著者らは以前より，データマイニング技術であるラフ集合を利用した橋梁データベースの解析を行い，ルール型知識の抽出を試みてきた．しかし，処理時間が膨大であり抽出した知識が評価されてなかった．従って，本研究では，これらの問題を解決するため，既存のシステムの改良を行い，抽出されたルールの検証を行った．さらにルール抽出に用いる属性の影響度を調べる方法を提案し，考察を行った．

2. ラフ集合によるルール抽出

2.1 ラフ集合の基本概念 [1)]

ラフ集合の基礎概念は，一言で言うと類別と近似である．われわれは外界からいくつかの情報に対して知的に行動するとき，それら情報における主語を属性に従って類別している．例えば橋梁を識別するとき，「橋齢」という属性だけに着目すれば，「橋齢」の等しいものは同じ物と類別される．この識別不能性がラフ集合の最も基本的な概念である．

2.2 決定表の簡約化

決定表の簡約化手順

ここで，本研究で用いた決定表簡約化の流れを図 1 に示す．

［Step1］**矛盾している決定規則を削除する**　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．

［Step2］**同一決定規則をまとめる**　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．

［Step3］**必要のない属性を削除する**　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．

［Step4］**必要のない属性を削除する**　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．

図 1　決定表簡約化の流れ

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．

3. 決定表簡約化手法の検証

3.1 既存のシステムにおける問題点の解決

本研究では問題を解決するためにシステムの高速化を考えた．本研究で構築したシステム（以後,本システム）では Access を使わず，配列を使うことで，計算コストの軽減を図った．実際に本年度に用いた橋梁の伸縮継手の異常音データ（条件属性 12 個，決定属性 1 個の決定表）を適用した結果，約 12 時間（既存のシステムの処理時間）に対し，約 5 分（本システムの処理時間）と，大幅に短縮できた．

4.2　決定表簡約化手法プログラムの検証

本システムを検証するため，fisher のあやめデータを用いた検証を行った[2]．検証方法としては，抽出されるルールの元データに対する識別率を求めることにした．fisher のあやめデータを度数分布から得たカテゴリー分けによりコード化した．コード化された fisher のあやめデータ 150 件の決定表を，本システムに適用した．

4.3　適用結果の考察

抽出されたルールによる fisher のあやめデータの識別数は 141 件と多く，また識別率も 94％とかなり高いことから，本システムで抽出されたルールによって正しくあやめの種類が類別されている．このことから，本システムは元データに対し，有用なルールを抽出できたといえる．

5.　実データへの適用

実際に扱われているデータを対象として，決定表簡約化手法に適用した．対象とするデータとして，阪神高速道路公団から頂いた道路高架橋の主桁に関する「資産データ」，「点検データ」を用いた．

5.1　主桁データへの決定表簡約化手法の適用

度数分布から主桁データを以下の 4 つのカテゴリーに分け，カテゴリー分けによるルール抽出の実験，項目選定によるルール抽出の実験を行った．

カテゴリー①：度数分布より，各属性値の分布しているところのまとまりを見て範囲を区切ったもの

カテゴリー②：①に対し，属性値のまとまりで範囲を区切るのではなく一定間隔で区切ったもの

カテゴリー③：度数分布より，属性値を区切る一定間隔を細かく取ったもの

カテゴリー④：度数分布より，属性値を区切る一定間隔を大きく取ったもの

5.2　カテゴリー分けによるルール抽出

カテゴリー①～カテゴリー④によってコード化された主桁データ 159 件の決定表を，本システムに適用する．これによって抽出されたルールについての矛盾レコード数，ルール数，識別数，識別率の調査について，4 つのカテゴリーの比較を行った．**表 1** に最も良い結果であった**カテゴリー④**の結果を示す．

表 1　カテゴリー④のルール抽出結

属性数	条件属性の部分集合	矛盾レコード数	ルール数	識別数	識別率
7	A03,A04,A05,A07,A09,A12,A15	121	102	122	76.73%
	A03,A04,A05,A07,A09,A12,A16		99	122	76.73%
	A03,A04,A05,A08,A09,A12,A15		89	125	78.62%
	A03,A04,A05,A08,A09,A12,A16		81	125	78.62%
8	A02,A03,A04,A05,A06,A08,A09,A15		68	118	74.21%
	A02,A03,A04,A05,A06,A08,A09,A16		62	118	74.21%
	A03,A04,A05,A06,A07,A08,A09,A15		78	113	71.07%
	A03,A04,A05,A06,A07,A08,A09,A16		72	113	71.07%
	A03,A04,A05,A06,A08,A09,A10,A15		81	127	79.87%
	A03,A04,A05,A06,A08,A09,A10,A16		73	127	79.87%
	A03,A04,A05,A06,A08,A09,A13,A15		68	118	74.21%
	A03,A04,A05,A06,A08,A09,A13,A16		62	118	74.21%
	A03,A04,A05,A06,A08,A09,A14,A15		68	118	74.21%
	A03,A04,A05,A06,A08,A09,A14,A16		62	118	74.21%
9	なし	—	—	—	—

5.3　項目選定によるルール抽出

属性を取り除くことでどうルール抽出に影響するかを考え，重要な項目を選定する．データとして，カテゴリー分けの結果より，一番良い結果と考えられる**カテゴリー④**の ***A03, A04, A05, A06, A08, A09, A10, A16*** の組合せを用いた．これによって抽出されたルールについての矛盾レコード数，ルール数，識別数，識別率の調査を行った．**表 2** に結果を示す．

表 2　項目選定によるルール抽出結果

条件属性の部分集合	矛盾レコード数	ルール数	識別数	識別率
A03,A04,A05,A06,A08,A09,A10,A16	126	72	118	74.21%
A04,A05,A06,A08,A09,A10,A16	133	34	86	54.09%
A03,A05,A06,A08,A09,A10,A16	135	51	95	59.75%
A03,A04,A06,A08,A09,A10,A16	138	28	95	59.75%
A03,A04,A05,A08,A09,A10,A16	128	63	105	60.04%
A03,A04,A05,A06,A09,A10,A16	130	62	109	68.55%
A03,A04,A05,A06,A08,A10,A16	128	56	105	66.04%
A03,A04,A05,A06,A08,A09,A16	136	38	106	66.67%
A03,A04,A05,A06,A08,A09,A10	128	65	116	72.96%

6.　まとめ

本研究で得られた成果を以下にまとめる．

① 既存のシステムの処理時間を大幅に短縮できた．

② 本研究で作成したシステムの決定表簡約化手法に対する有効性が示された．

③ 抽出されたルールの各属性の影響度が分かり，重要な項目が把握できた．

参考文献

1）中村　昭：ラフ集合－その基本概念と知識情報，数理科学 No373，サイエンス社，1994.7

2）http://sun.econ.seikei.ac.jp/~shinmura/archive.html