様式 C-19

科学研究費補助金研究成果報告書

平成２１年４月１３日現在

研究種目：基盤研究（Ｃ）
研究期間：2006～2008
課題番号：１８５００２１４
研究課題名（和文）多様なデータも処理するラフデータマイニングツールの構築と応用

研究課題名（英文）A Generation of a Rough Sets Based Data Mining Tool for a Wide Variety of Data and Its Application

研究代表者
酒井　浩（SAKAI HIROSHI）
九州工業大学･工学研究院･教授
研究者番号：６０２０１５１３

研究成果の概要：本研究では、ラフ集合の特長を生かしてデータマイニングツールを構築し、ツールを多様な表データ（広くはデータベース）に適用し、価値ある情報（具体的には相関ルールと呼ばれる含意式「ある条件が成立する場合に、特定の状況が恒常的に起きる」）の獲得支援を行った。特に、大規模な場合分けが生じる非決定情報表に対して従来のアプリオリアルゴリズムを拡張し、場合分けの数が10の100乗を超えるような情報の不完全性を有する非決定情報表からも相関ルールの取り出しが瞬時にできることを確認した。本手法は不完全情報に基づくデータマイニングの新たな枠組みになると考えられる。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,200,000	0	1,200,000
2007 年度	1,100,000	330,000	1,430,000
2008 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,400,000	660,000	4,060,000

研究分野：総合領域
科研費の分科・細目：情報学･統計科学
キーワード：ラフ集合、データマイニング、相関ルール、回帰直線、非決定情報、アプリオリアルゴリズム、粒状計算、数値パターン

１．研究開始当初の背景

(1) タイトルにある「ラフデータマイニング」は、ラフ集合理論に基づくデータマイニングを意味する。本研究の目的は、ラフ集合の特長を生かしてデータマイニングツールを構築し、ツールを多様な表データ（広くはデータベース）に適用し、価値ある情報の獲得支援を行うことである。ラフ集合では通常、属性値が離散値である表データを研究対象とするが、それ以外の多様なデータ、例えば、属性値が連続値である表データ、属性値に情報の欠落・不完全性がある表データ、広くは表データ化されていない無構造データ、などに対するラフデータマイニングは今後の重要な研究課題になると考えた。

(2) 統計科学の分野では既に多変量解析としてデータ解析手法が確立しており、重回帰分析などによりデータの特徴を把握できる。しかし、多変量解析は平均や分散を定義できないカテゴリカルデータの処理には向かない。ラフ集合では逆に、カテゴリカルデータから自然に定義される同値関係を利用するので、多変量解析の弱点を補う。例えば、本学における入試のアンケートから抽出された相関ルール「就職に有利という理由で本学を受験した学生の殆どは学部卒業後に就職を希望する」は多変量解析では処理しにくいが、ラフ集合の性質により容易に取り出すことができる。相関ルールは多変量解析における一種の回帰直線と考えられ、上記の相関ルールは大学院志望者数の推定に利用できる。このように、ラフ集合と多変量解析の利点を生かしてデータマイニング機能の強化ができると考えた。ラフ集合と多変量解析を融合した新たなデータ解析法への発展も期待できると考えた。

(3) 表における連続値データの処理では、通常のラフ集合による類別を考えると類の個数が多くなり過ぎる欠点があった。例えば、身長について 175.8cm のように少数第１位まで見れば、乳児の 30.0cm から大人の 200.0cm までおよそ 1700 種類の類が存在する。1700 個の類はラフ集合で扱うには多すぎる。もちろん 1700 個の類を計算機上で処理することは可能であるが、得られる含意式は頻度の面で極めて小さく、一般に相関ルールと呼ぶには難点があると考えられる。このような側面にラフ集合の枠組みをどのように適用するか検討する必要性があった。

(4) 情報の欠落・不完全性がある表データの処理では、申請時点でかなりのプログラムを実現しており、アンケートデータなどの解析を行いながらプログラムの改善を進める計画であった。さらに、不完全性を処理するために相関ルールに様相性を取り込む研究を継続した。研究代表者は、表にあいまいデータを記述する枠組み「非決定情報システム」における処理アルゴリズムの殆どを既に明らかにしており、従来手付かずにされていた処理も実現可能になりつつあった。

(5) その他、構築したツールの応用の場として、UCI 機械学習レポジトリのデータにおけるマイニングについても検討を始めていた。

２．研究の目的

(1) 本研究では、次の２課題を考慮し、理論的枠組みと実際のツールの実現・応用を図ることを目指した。

課題１：『連続値データのためのラフ集合処理と実データへの応用』

課題２：『情報の欠落・不完全性がある表データのためのラフ集合処理と実データへの応用』

(2) 課題１：『連続値データのためのラフ集合処理と実データへの応用』では、次の３点を検討した。

① 最も注目したものは数値データの桁数による離散化である。例えば、円周率πは無理数であり小学生には３程度の数で十分であるが、数式処理の研究者にとっては 3.1415 でも不十分である。数字の桁に注目することで詳しく見た場合の同値類、おおざっぱに見た場合の同値類を自然に定義でき、詳しく見た場合の相関ルール、おおざっぱに見た場合の相関ルールへと発展できると考えた。

② 野球のデータで「打率⇒打点」で条件部が「打率=２割５分」の相関ルールがあるとして、打率を２割代（少数以下第１位のみ見る）の条件まで広げるとサポート（指標値の１つで、全データに対する本含意式の出現率）は増加し、正確度（指標値の１つで、条件部の出現回数に対する本含意式の出現率、これで矛盾の程度を把握する）は下がる。逆の操作では、サポートは下がり、正確度は上がる。注目する桁数を制御しながら最も妥当な相関ルールの抽出を行うことが可能になると考えた。

③ 上記の内容を陽に扱うために数値パターンの導入を検討し、UCI 機械学習レポジトリのデータを中心に作成したツールによる具体的な相関ルールの取り出しを目指した。

(3) 課題２：『情報の欠落・不完全性がある表データのためのラフ集合処理と実データへの応用』では、不完全情報表の欠落部分を起こり得る値で置き換えた決定情報表（派生する決定情報表と呼ぶ）に基づいた様相的な処理法を検討する。派生する決定情報表は通常、指数関数的に増加するためすべての派生する決定情報表を列挙することは難しい。この原因から、様相性に依存した処理は手付かずのまま残っている。この点をラフ集合の枠組みにより改善することを目指した。

３．研究の方法

決定情報表における相関ルール抽出問題は既に確立されており、通常、下記の問題として定式化されることが多い。

【決定情報表における相関ルール抽出問題】
２つの閾値をαとβとする。決定情報表に出現する含意式τで以下の条件を満たすものを相関ルールの候補として全て取り出せ。
(A) τのサポートがα以上である。
(B) τの正確度がβ以上である。

直観的には、一定の頻度以上起きており、かつ矛盾が少ない含意式τを取り出すものである。本問題を解決するアルゴリズムとして、アグラワル（Agrawal）によるアプリオリアルゴリズムが有名である。研究方法として、本アルゴリズムをそれぞれのデータ用に拡張することを考えた。また、種々のデータに応用することを目指した。

４．研究成果
(1) 課題１：『連続値データのためのラフ集合処理と実際のデータへの応用』における成果を列挙する。

① 注目する桁を記号@、注目しない桁を記号#で記述し@と#の文字の列として数値パターンを定義した。数値パターンにおける@の数値が同じ２つの数 A と B は同じ値とみなす。@@@では身長 172cm と 175cm は異なるが、@@#では２つの値は身長 170cm 代となり同じ数とみなす。このような数値パターンを使うラフ集合の理論を構築した。

② 連続値データのための相関ルール抽出問題を以下のように定義した。

【連続値データでの相関ルール抽出問題】
２つの閾値をαとβとする。決定情報表に出現する数値パターン付き含意式τで以下の条件を満たすものを相関ルールの候補として全て取り出せ。
(A) τのサポートがα以上である。
(B) τの正確度がβ以上である。

本問題に対して記号@と#も処理するアプリオリアルゴリズムを実現した。C#言語によりユーザインターフェースも実現している。

③ 実際に UCI 機械学習レポジトリにおける肝炎データ Hepatitis.csv(対象数 80、全データ数は 155 であるが欠落値を含まない 80 個を対象とした。属性数 20)に対して回帰分析を行い

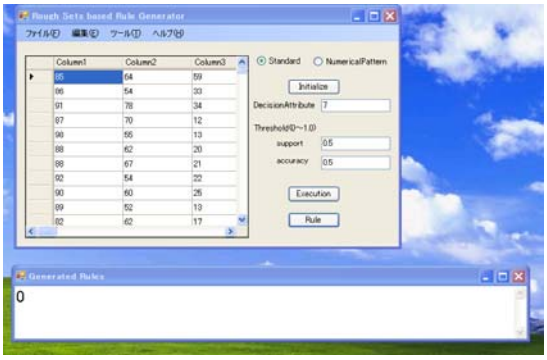
患者の生死＝0.227×体調不良－0.390×食欲不振＋0.186×腹水－0.068×総ビリルビン値＋0.002×プロトロンビン時間－0.189×肝生検＋2.05

の回帰式を得た。しかし、本肝炎データに対して数値パターンを用いた手法により

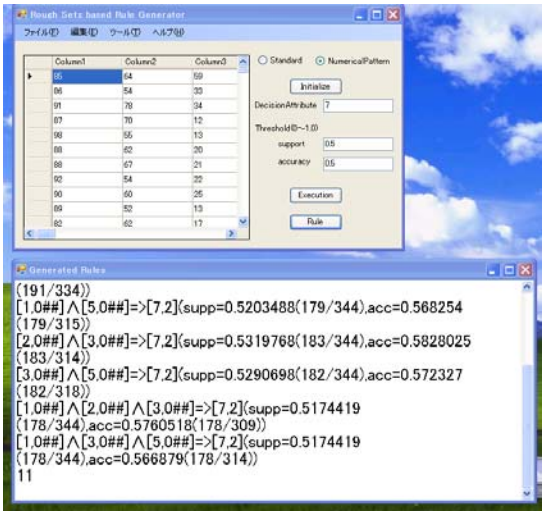
相関ルール：「静脈瘤がある」かつ「アルブミンが４以上５未満」ならば「患者は生存している」

の取り出しができた（数値パターンの利用がなければ、アルブミンが４以上５未満といった区間値の処理に難点が生じ、このような相関ルールの取り出しはできない）。このパターンは全体の約半数で矛盾なく起こっていた。このようにして、連続値データからも効率よく相関ルールが取り出せた。

④ 下記の２つの図は UCI 機械学習レポジトリにある肝臓病のデータ Bunpa.csv（対象数 345、属性数 7）の実行例である。条件属性は数値データから成るので、通常のラフ集合（Standard ボタンの選択）ではサポート 0.5 以上を満たす相関ルールは存在しない。



しかし、本データに対して数値パターンの選択（NumericalPattern ボタンの選択）を行うと、サポート 0.5 以上かつ正確度 0.5 以上を満たす 11 個の相関ルールを得ることができた。図における最下位の数 11 は、#を含む 11 個の相関ルールの存在を示す。



このように、数値データを適度に調整する機能をアプリオリアルゴリズムに追加し、従来、相関ルールとして取り出し得なかった含意式の抽出を可能にできた。

⑤ 従来のラフ集合における数値データの処理では、エントロピーなどの散らばり指標により数値データ自体の離散化が行なわれ、その後に離散化されたデータにラフ集合の手法が適用された。本手法では離散化の手続きは不要であり、ルールで出現する数値の階層性を利用するので、直感的でわかり易い相関ルールを抽出できる。

⑥ 以上に示したように、提案する枠組みは数値データに対する新たな取り組みであり、従来の多変量解析などとの関連から今後も重要な研究課題になると考える。

(2) 課題２：『情報の欠落・不完全性がある表データのためのラフ集合処理と実際のデータへの応用』における成果を列挙する。

① 研究代表者は派生する決定情報表に依存した相関ルール抽出問題を以下のように定義した。

【非決定情報表での相関ルール抽出問題】
２つの閾値をαとβとする。
［最悪系の相関ルール抽出］
　全ての派生する決定情報表において、以下の条件を満たすものを相関ルールの候補として全て取り出せ。
　(A) τのサポートがα以上である。
　(B) τの正確度がβ以上である。
［最良系の相関ルール抽出］
　派生するある決定情報表において、以下の条件を満たすものを相関ルールの候補として全て取り出せ。
　(A) τのサポートがα以上である。
　(B) τの正確度がβ以上である。

最悪系では派生するすべての決定情報表で条件が成立するために、得られた相関ルールは情報の不完全性を受けず非決定情報表においても常に成立すること(Certainty)が保証される。一方、最良系では含意式にとって最も都合の良い決定情報表で条件が満たされればよいので、可能性(Possibility)を考慮した相関ルール抽出問題になっている。

② 上記、２つの系の定義は可能であるが、実際に派生する決定情報表を列挙する手法では指数オーダーの問題が生じる。我々は本問題をラフ集合における粒状計算によって下記のように解決した。

［結果１］
派生する決定情報表には、サポートと正確度を共に最小にする表がある。また、共に最大にする表もある。
［結果２］
含意式τのサポートと正確度の最小値を計算する式がある。また、含意式τのサポートと正確度の最大値を計算する式がある。これらの計算式は派生する決定情報表の個数には依存しない。
［結果３］
上記の性質を利用し、最悪系の問題は指標値の最小値を利用するアプリオリアルゴリズムで処理できる。最良系の問題は指標値の最大値を利用するアプリオリアルゴリズムで処理できる。
［結果４］
非決定情報表における最悪系と最良系の計算時間は決定情報表におけるアプリオリアルゴリズムの処理時間と殆ど変わらない。

上記の結果から、提案する相関ルール抽出問題は不完全な情報まで処理できてしかも決定情報表を処理する時間とあまり変わらない。派生する決定情報表の個数に依存しない処理になっており、実用面でも十分利用できる枠組みであると考える。現在、C 言語により 2000 行程度のプロトタイプ版を実現している。

③ 研究代表者は非決定情報表での相関ルール抽出問題を不完全な表データに適用した。下記の表は UCI 機械学習レポジトリにある乳がんデータ Mammographic.csv（対象数 150、属性数 6）の先頭の 5 個分を EXCEL に取り込んだものである。この中には 5 個の欠落値？が存在する。

4	40	1	1	?	1
4	20	1	1	3	0
5	70	1	5	?	1
4	60	1	?	3	0
4	70	?	?	3	0

この？の取る値が有限個（4 個または 5 個）であることを利用して、本情報表を下記の非決定情報表に書き換え、不完全情報表をラフ集合非決定情報解析の問題に転換する。

4	40	1	1	[1,2,3,4]	1
4	20	1	1	3	0
5	70	1	5	[1,2,3,4]	1
4	60	1	[1,2,3,4,5]	3	0
4	70	[1,2,3,4]	[1,2,3,4,5]	3	0

実際、実験に用いた 150 対象については、76 個の？記号が存在し $4^{55} \times 5^{21}$ 通りの派生する決定情報表が存在したが、下記のように実時間で実行した。下記は最悪系の実行の様子を示す。

```
sakai@vaio /cygdrive/m/ms35saka(Sep1108)/m-mammo
$ ../sanisapri
version 1.2.8
File Name: mammo
==================================
 Lower Approximation Strategy
==================================
CAN(1)={[AGE,40],[AGE,50],[AGE,60],[SHAPE,1],
[SHAPE,2],[SHAPE,4],[MARGIN,1],[MARGIN,4],
[DENSITY,3],[SEVERITY,0],[SEVERITY,1]}(11)

CAN(2)={[SHAPE,1][SEVERITY,0](<DEF>0.729,
<INDEF>0.735),[SHAPE,2][SEVERITY,0](<DEF>0.750,
<INDEF>0.756),[MARGIN,1][SEVERITY,0](<DEF>0.803,
<INDEF>0.806),[DENSITY,3][SEVERITY,0](<DEF>0.375,
<INDEF>0.382),[SHAPE,4][SEVERITY,1](<DEF>0.712,
<INDEF>0.717),[DENSITY,3][SEVERITY,1](<DEF>0.391,
<INDEF>0.397)}(6)

########## OBTAINED RULE ##########
[SHAPE,1]=>[SEVERITY,0](minsupp<DEF>=0.233,
minsupp<INDEF>=0.240,minacc<DEF>=0.729,
minacc<INDEF>=0.735)
(<DEF>from4,6,8,13,16,20,23,27,31,34,35,42,43,48,
53,67,76,78,86,88,89,93,95,97,105,106,109,116,122,
124,128,139,143,144,150)
(<INDEF>from 7,49,84,129)
[SHAPE,2]=>[SEVERITY,0](minsupp<DEF>=0.200,
minsupp<INDEF>=0.207,minacc<DEF>=0.750,
minacc<INDEF>=0.756)
          :     :     :
EXEC_TIME=0.015(sec)
```

その他、UCI 機械学習レポジトリにおける肝炎のデータ Hepatitis.csv （対象数 155、属性数 20、欠落値 167）については起こり得る場合の数が 10 の 100 乗を超えたが、含意式

[MALAISE,2]=>[Class,2]

[SPIDERS,2]=>[Class,2]

は起こり得る全ての場合においてサポートが 0.5 以上かつ正確度が 0.9 以上を満たす相関ルールとして取り出すことができた。

④ 非決定情報表での相関ルール抽出問題については従来にない手法として、ラフ集合の国際会議 RSCTC08 で注目された。最新の結果は、発表論文①にまとめてある。さらに、本研究を高く評価したカナダインフォブライト社の研究者と共に、今回の研究を汎用の MySQL 上に実現する共同研究を最近開始した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 18 件）

① H. Sakai, R. Ishibashi, M. Nakata, Rules and Apriori Algorithm in Non-deterministic Information Systems, Transactions on Rough Sets, Vol.9, 328-350, 2008, 査読有

② H. Sakai, K. Koba, M. Nakata, Rough Sets Based Rule Generation from Data with Categorical and Numerical Values, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.12, 426-434, 2008, 査読有

③ H. Sakai, R. Ishibashi, M. Nakata, Lower and Upper Approximations of Rules in Non-deterministic Information Systems, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.5306, 299-309, 2008, 査読有

④ M. Nakata, H. Sakai, Applying Rough Sets to Information Tables Containing Possibilistic Values, Transactions on Computational Science, Vol.2, 180-204, 2008, 査読有

⑤ M. Nakata, H. Sakai, Rough Sets Approximations in Data Tables Containing Missing Values, Proc. IEEE World Congress on Computational Intelligence FS0165, 2008, 査読有

⑥ H. Sakai, M. Nakata, On Possible Rules and Apriori Algorithm in Non-deterministic Information Systems 2, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.4482, 280-288, 2007, 査読有

⑦ H. Sakai, K. Koba, R. Ishibashi, M. Nakata, On a Rough Sets Based Tool for Generating Rules from Data with Categorical and Numerical Values, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.4617, 269-281, 2007, 査読有

⑧ M. Nakata, H. Sakai, Applying Rough Sets to Information Tables Containing Probabilistic Values, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.4617, 282-294, 2007, 査読有

⑨ M. Nakata, H. Sakai, Lower and Upper Approximations in Data Tables Containing Possibilistic Information, Transaction on Rough Sets, Springer-Verlag, Vol.7, 170-189, 2007, 査読有

⑩ M. Nakata, H. Sakai, Applying Rough Sets to Data Tables Containing Missing Values, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.4585, 181-191, 2007, 査読有

⑪ M. Nakata, H. Sakai, Rough Sets Dealing with Data Tables Containing Missing Values, The International Conference on Soft Computing, Intelligent Systems and Information Technology, 147-153, 2007, 査読有

⑫ H. Sakai, On a Rough Sets based Data Mining Tool in Prolog: An Overview, Lecture Notes in Artificial Intelligence (Selected

Papers), Springer-Verlag, Vol.4369, 48-65, 2006, 査読有
⑬ H.Sakai, M.Nakata, An Application of Discernibility Functions to Generating Minimal Rules in Non-deterministic Information Systems, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.10, 695-702, 2006, 査読有
⑭ H.Sakai, M.Nakata, On Possible Rules and Apriori Algorithm in Non-deterministic Information Systems, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.4259, 264-273, 2006, 査読有
⑮ M.Nakata, H.Sakai, Applying Rough Sets to Data Tables Containing Imprecise Information under Probabilistic Interpretation, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.4259,213-223, 2006, 査読有
⑯ H.Sakai, K.Koba, M.Nakata, An Application of Rough Non-deterministic Information Analysis to Class Evaluation Data by Students, Kansei Engineering International, Vol.6, 25-32, 2006, 査読有
⑰ M.Nakata, H.Sakai, Applying Rough Sets to Data Tables Containing Possibilistic Information, Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol.4062, 147-155, 2006, 査読有
⑱ M.Nakata, H.Sakai, Rough Sets Approximations to Possibilistic Information, Proc. IEEE World Congress on Computational Intelligence, No.4381, 1-8, 2006, 査読有

〔学会発表〕(計１０件)
① 林康平、中田典規、酒井浩、粒状計算と表データにおける制約充足問題について、TE2-2、第24回ファジィシステムシンポジウム, 2008年９月４日、大阪、阪南大学
② 酒井浩,林康平、中田典規、非決定情報表におけるルールの上近似と下近似について、TE2-3、第24回ファジィシステムシンポジウム, 2008年９月４日、大阪、阪南大学
③ 中田典規、酒井浩、欠損値を含む情報テーブルにおけるラフ近似、WE3-1、第24回ファジィシステムシンポジウム, 2008年９月３日、大阪、阪南大学
④ 酒井浩、木場和博、中田典規、連続値連続値データも処理するラフデータ解析ツールについて、第４回日本感性工学会、2008年３月７日、仙台、宮城大学
⑤ 高原祐起、中田典規、酒井浩、ラフ集合とアプリオリ, 多変量解析, 決定木などの手法によるデータ解析の考察、第５回ラフ集合と感性工学ワークショップ、35-38、2007年６月３日、東京、東海大学
⑥ 木場和博、中田典規、酒井浩、連続値データにおける数値パターンの導入とラフ集合に基づくルール生成について、第 23 回ファジィシステムシンポジウム, 2007 年８月３０日、名古屋、名城大学
⑦ H.Sakai, K.Koba, M.Nakata, On Issues in Rough Non-deterministic Information Analysis (RNIA), Proc. 9th Czech-Japan Seminar, 65-70, 2006 年８月１９日、北九州、早稲田大学
⑧ M.Nakata, H.Sakai, 可能性分布で表された情報を含むデータへのラフ集合の適用, 第 22 回ファジィシステムシンポジウム, 681-684, 2006 年９月１５日、北海道、北海学園大学
⑨ H.Sakai, K.Koba, M.Nakata, On Rough Sets Based Rule Generation from Tables with Numerical Values, Proc. SCIS & ISIS2006, pp.SU-D2-2, 1-6, 2006 年９月２１日、東京、東工大
⑩ M.Nakata, H.Sakai, Rough Approximations under Two Interpretations of Missing Values, Proc. SCIS & ISIS2006, pp.SU-D2-4, 1-6, 2006年９月２１日、東京、東工大

〔図書〕(計１件)
① H.Sakai, Rough Non-deterministic Information Analysis and Related Issues, Handbook on Reasoning-based Intelligent Systems (Accepted), World Scientific, 1-35, 2008, 査読有

６．研究組織
(1)研究代表者
酒井　浩　(SAKAI HIROSHI)
九州工業大学･工学研究院･教授
研究者番号：６０２０１５１３

(2)研究分担者

(3)連携研究者

(4)研究協力者
中田　典規　(NAKATA MICHINORI)
城西国際大学･経営情報学部･教授
研究者番号：１０２０１６６７