様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年６月１０日現在

機関番号：１０１０３
研究種目：若手研究（Ｂ）
研究期間：2011 ～ 2012
課題番号：２３７００２４４

研究課題名（和文）ランダムラフ集合による大規模感性データマイニングシステムの構築

研究課題名（英文）Development of a Kansei data mining system based on random rough sets

研究代表者
工藤　康生　（KUDO YASUO）
室蘭工業大学・工学研究科・准教授
研究者番号：　９０３６０９６６

研究成果の概要（和文）：本研究は，大規模データに対するラフ集合を用いたデータマイニングの実現を目的とし，まず統計的集団学習の手法をラフ集合に導入したランダムラフ集合モデルを提案し，その基礎理論の構築を行った．更にそれに基づくデータマイニングシステムを実装し，その有効性を実験により検証した．その結果，大規模データに対するランダムラフ集合に基づくデータマイニングについて，基盤技術を確立することができたと考える．

研究成果の概要（英文）：In this study, we proposed a random rough set model by introducing a concept of ensemble learning to rough sets. We also developed a Kansei data mining system based on the random rough set model and evaluated the developed system by experiments. Consequently, our study contributed the development of essential basis of rough set-based data mining for Kansei data with many samples and attributes.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,400,000	720,000	3,120,000

研究分野：総合領域
科研費の分科・細目：情報学，感性情報学・ソフトコンピューティング
キーワード：あいまいと感性

１．研究開始当初の背景

ラフ集合を用いたアンケートデータ等の分析手法は，データマイニング技術の感性情報処理への応用として近年注目を集めている．ラフ集合によるデータ分析は主に，対象となるデータを正しく分類するために最小限必要となる属性を抽出する縮約計算，およびデータ内の規則性を記述する決定ルールの抽出に基づく．しかし，これらの計算コストはデータの規模の増大に対して指数的に増加するため，ラフ集合によるデータ分析は比較的小規模なデータを対象とせざるを得ず，購買記録や Web 上でのアンケート調査結果などの，大規模データに対するラフ集合の適用は現実的には困難であった．

２．研究の目的

本研究は，感性工学の分野で注目されているラフ集合を用いて，購買履歴等，ユーザの感性を反映した大規模データに対する分析手法を構築することを目的とする．特に本研究では，大規模データに対する統計的な集団学習の手法をラフ集合に導入したランダムラフ集合モデルを提案することにより，大規模データからの大域的な傾向の抽出と大規模データに潜む局所的な特徴の抽出を両立させることに焦点を当てる．

３．研究の方法

本研究では，ラフ集合によるデータ分析と統計的集団学習の手法及びメリット，デメリットに着目する．

ラフ集合によるデータ分析：
- 手法：決定表として与えられたデータに対してすべての縮約及び極小決定ルールを抽出.
- メリット：すべての極小決定ルールを抽出することにより，局所的な特徴も抽出可能.
- デメリット：計算量が膨大なため，大規模データに対する使用が困難.

統計的集団学習：
- 手法：リサンプリングにより複数小規模データを生成し，各データに対する学習により得られた結果を統合.
- メリット：アルゴリズムが比較的高速であり大規模データに対しても適用可能.
- デメリット：統計的な学習手法であるため局所的な特徴が抽出できない恐れがある.

本研究ではこれらのメリットおよびデメリットを踏まえ，以下の（１）～（３）の研究を行う.

（１）統計的集団学習の手法をラフ集合に導入することにより，1)大規模データからリサンプリングにより複数小規模のデータ（決定表）を作成，2)各決定表に対して縮約計算及び極小決定ルールを抽出，3)抽出結果を分析・統合することにより，大規模データ全体に対する縮約計算及び極小決定ルールの抽出を近似的に行う手法（ランダムラフ集合）の数理モデルを構築する.

（２）構築したランダムラフ集合の数理モデルを計算機上で実装し，評価実験を通じてその有効性を検証する．特に，リサンプリングにより作成する小規模決定表の個数，小規模決定表におけるサンプル数，属性数などのパラメータはランダムラフ集合によるデータ分析に大きな影響を与えるため，詳細な評価実験を行うことにより，適切なパラメータを見出す.

（３）実装したランダムラフ集合モデルを用いて，複数小規模決定表に対する縮約計算及び極小決定ルール抽出による局所的な特徴抽出と，その結果を分析・統合することによる大域的な傾向抽出の機能を併せ持つ，大規模感性データに対するデータマイニングシステムを実装し，評価実験を通じてその有効性を検証する.

４．研究成果

（１）研究の主な成果

①統計的集団学習の手法をラフ集合に導入したランダムラフ集合モデルを構築した（学会発表[1],[3]）．構築したモデルは，Bazan et al.による縮約計算の統計的アプローチである Generalized Dynamic Reduct (GDR)と，研究代表者らの従来研究である，属性の個数が多いデータから小規模決定表を生成することによりできるだけ多数の縮約を抽出するヒューリスティックな手法とを併用しており，以下の手順に沿って，サンプル数および属性の個数が共に多いデータに対する縮約計算を近似的に行う.

1. 大規模データに対してリサンプリングを行い，データ数が比較的少数の決定表（部分表）を多数生成する.
2. 生成した各部分表に対して，サンプルの分類能力を保持したまま属性の個数を削減した小規模決定表を多数生成する.
3. 小規模決定表に対して縮約計算を行い，大規模データ全体に対する縮約の候補を多数抽出する.
4. 分類能力チェック用の部分表を多数生成し，3.で生成した各縮約候補の分類能力を検証する．一定の割合以上のチェック用小規模決定表でデータを正しく分類できた縮約候補を，大規模データ全体に対する縮約計算の結果として出力する.

手順 2.で各部分表から小規模決定表を多数生成する手法は，研究代表者の従来研究による知見を用いている．また，手順 3.で小規模決定表から抽出した縮約は，小規模決定表の基となった部分表の縮約となることが理論的に保障されている．そのため，手順 3.で抽出した縮約は，GDR の知見を用いることにより，手順 4.であらかじめ設定した，分類に失敗する決定表の割合の許容限度を表すしきい値 $\varepsilon(0 \le \varepsilon < 0.5)$ に基づいて，多数生成したチェック用部分表の $(1-\varepsilon)\times 100\%$ 以上でデータを正しく分類できれば，その縮約は大規模データ全体に対する近似的な縮約（ε-GDR）と見なすことができる.

提案したランダムラフ集合モデルを計算用サーバ上で実装し，UCI ML Repository の 2 種類のベンチマークデータ（Internet Advertisement dataset, CANE-9 dataset）に対して ε-GDR を抽出する実験を行った．また，上記の手順 1.で生成する各部分表に含まれるサンプルの個数および許容限度 ε を変化させ，これらのパラメータが結果に与える影響を調査した．IA dataset での結果の例を表 1.に示す（学会発表[1]）.

表 1. Internet Advertisement dataset に対する ε-GDR の抽出結果

size	ε-GDR			
	$\varepsilon = 0.3$	0.2	0.1	0.05
20%	3002	1013	101	1
30%	5708	1880	208	0
40%	8493	5401	1955	176
50%	9743	9108	6064	2053

表 1.で，項目 size は手順 1.で生成した各部分表の，IA dataset 全体に対するサンプル数の割合を表しており，データ全体の 20%から 50%までの 4 段階とした．許容限度は $\varepsilon = 0.3$ から 0.05 までの 4 段階とした．また，この実験では部分表の 4 段階のサイズそれぞれについて，手順 3.で縮約の候補を約 10000 個抽出し，手順 4.では確認用部分表の個数を 300 個とした．よって，例えば size=20%，$\varepsilon = 0.3$ の値である 3002 は，今回の実験では，「データ全体の 20%の大きさの部分表を多数生成し，これらから得られた縮約の候補の中で，3002 個の候補は，手順 4.で生成した 300 個の確認用部分表の 70%(=210 個)以上でデータを正しく分類できたため，ε-GDR と見なされる」ことを表す．

実験の結果，手順 1.で生成する各部分表のサイズが大きいほど，許容限度をより小さく設定した ε-GDR を多数抽出できる傾向が見られた．この傾向は CANE-9 dataset でも同様であった．許容限度を小さく設定した ε-GDR ほど，未知のデータに対する頑健性が高いと見なすことができるため，データ全体の近似的な縮約として適切であると考えられる．一方，個々の部分表のサイズが大きくなるほど，縮約抽出に要する時間も増大するため，部分表のサイズと許容限度とのバランスを検討する必要がある．この問題は今後の課題とする．

また，ランダムラフ集合モデルの構築に関連して，決定表からのルール抽出の際に使用する属性値の個数をできるだけ削減する値縮約の手法の改良（雑誌論文[2]）も行った．

②ランダムラフ集合モデルを用いて縮約計算を近似的に行う際に，縮約計算を並列分散化することによる計算の高速化は，本研究の目的である大規模データに対するラフ集合データマイニングの実現に向けた重要な課題の一つである．この課題に対して，研究代表者らの従来研究である，属性の個数が多いデータから小規模決定表を生成することによりできるだけ多数の縮約を抽出するヒューリスティックな手法を並列分散化し，縮約計算を高速に実行することを可能にした（雑誌論文[1]，学会発表[4],[6]）．

提案した高速化手法では，元となるヒューリスティックな縮約生成法であらかじめ定めた個数まで小規模決定表の生成を繰り返す際に，それぞれの小規模決定表の生成には他の小規模決定表に関する情報を一切用いないことに着目し，各小規模決定表の生成および縮約計算のプロセスを並行して実行することにより，小規模決定表の生成および得られた小規模決定表からの縮約計算を並列分散化している．得られた小規模決定表は十分に小さい決定表であるため，これに対するすべての縮約の抽出は高速に行うことができる．よって，提案した高速化手法では，属性の個数が多いデータから多数の縮約を高速に抽出することが可能である．

この高速化手法を計算用サーバ上で実装した．実装に際して言語は C++を用い，並列分散化には Open MP を用いた．使用した計算用サーバは Linux サーバ（CPU: intel Xeon X5690 6Core 3.46GHz×2，メモリ:96GB，HDD: 1TB，OS: Cent OS 4.5）であり，最大で 12 スレッドを並列に実行可能である．

提案手法による高速化の実用性を検証するため，上述の環境で提案手法を 5 種類のベンチマークデータ（Audiology，Breast，IA，Leukemia，Lung cancer）に対して使用した．各データセットに対して，用いるコアの個数を 1（並列化しない逐次計算）～12 とした場合の計算時間をそれぞれ計測した．実験は 50 回行い，各データセットの各コア数について平均計算時間を求めた．実験結果を表 1.に示す（雑誌論文[1]）．

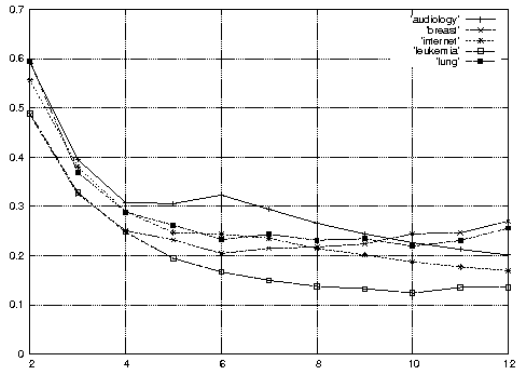

図 1. 並列分散化の実験結果

図 1.の横軸は計算に用いたコアの個数を，縦軸は各データセットについて，並列化しない逐次計算に要した時間に対する，各コア数での計算時間の割合を表す．いずれのデータセットでも，使用するコア数を 1 から 2 に増やすことで，計算時間は逐次計算に要した時間の 50%～60%に減少した．一般的に，使用するコアの個数が増加するほど計算時間は短くなる傾向が見られ，計算時間の減少の効果が最も顕著に表れた Leukemia dataset では，コア数 10 の場合に，逐次計算に要した時間の約 12%まで計算時間を短縮できた．

一方，コア数の増加に対して計算時間の減少率は低下する傾向が見られた．これは，今回実装した並列化ではループ処理のみの並列化であり，ループ内で実行する小規模決定表の作成および縮約計算に要する計算時間は短縮されていないためと考えられる．アルゴリズムの改良による小規模決定表の生成の高速化，および使用するコア数の調整など

については，今後の課題とする．
縮約計算の並列分散化と関連して，並列分散環境におけるユーザのログデータからのマイニング手法についても研究を行った（学会発表[2]）．

（２）得られた成果の国内外における位置づけおよびインパクト
研究期間全体を通しての成果として，ランダムラフ集合の基礎理論の構築，および縮約抽出手法の並列分散化，ランダムラフ集合に基づくデータマイニング手法の実装を行った．その結果，大規模データに対するランダムラフ集合に基づくデータマイニングについて，基盤技術を確立することができたと考える．
研究開始当初の背景で述べた通り，大規模データに対するラフ集合の適用は現実的には困難であると従来は考えられていた．これに対し，本研究による成果を用いることで，購買記録や Web 上でのアンケート調査結果などの大規模データに対しても，ラフ集合によるデータ分析が現実的に行えると考えている．よって，本研究の成果は，ラフ集合理論の感性情報処理への応用面に大きな進展をもたらすと期待できる．

（３）今後の展望
大規模データに対するランダムラフ集合に基づくデータマイニングの基盤技術を用いて，従来のラフ集合データ分析では困難であった，２つの属性間の比較に基づく特徴の抽出を行うことに基づく，ラフ集合による関係性マイニング手法の確立を目指す．ラフ集合による関係性マイニングは，例えば「商品Ａより商品Ｂをより好ましいと評価している」など，2 つの項目間の関係性に基づく特徴の抽出を目的とする．基礎的なアイデアは，決定ルールの値縮約に関する研究（雑誌論文[2]）および 2 つの属性間の関係性を新たな属性として表現する手法（学会発表[5]）として発表済みである．今後はこれらのアイデアを更に発展させ，具体的なアルゴリズムの構築および実装，実験による検証などを行う予定である（ラフ集合による関係性マイニング―感性データ分析の新展開―，平成 25 年度基盤研究（Ｃ），課題番号 25330315）．

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 2 件）
[1] Yasuo Kudo and Tetsuya Murai, A Parallel Computation Method for Heuristic Attribute Reduction Using Reduced Decision Tables, JACIII, 査読有, Vol.17, No.3, pp. 371-376, 2013.

[2] Yasuo Kudo, A Revised Approach to Solving the Symbolic Value Partition Problem from a Viewpoint of Roughness of Partitions, Int. J. Reasoning-based Intelligent Systems, 査読有, Vol.4, No.5, pp. 129-139, 2012.

〔学会発表〕（計 7 件）
[1] Yasuo Kudo and Tetsuya Murai, An Attempt of Hybridization of Generalized Dynamic Reducts and A Heuristic Attribute Reduction Using Reduced Decision Tables, FUZZ-IEEE 2013, 2013 年 7 月 8 日～10 日（発表確定），ハイデラバード（インド）．

[2] S. K. Shrestha, Y. Kudo, B. P. Gautam, and D. Shrestha, Multidimensional Service Weight Sequence Mining based on Cloud Service Utilization in Jyaguchi, IMECS 2013, 2013 年 3 月 13 日～15 日，香港(中国)．

[3] 工藤康生，村井哲也，ラフ集合および統計的手法に基づく大規模データからの縮約抽出について，第 28 回ファジィシステムシンポジウム，2012 年 9 月 12 日～14 日，名古屋（日本）．

[4] Yasuo Kudo and Tetsuya Murai, A Parallel Computation Method of Attribute Reduction, ISCIIA 2012, 2012 年 8 月 20 日～23 日，札幌（日本）．

[5] Yasuo Kudo and Tetsuya Murai, Indiscernibility Relations by Interrelationships between Attributes in Rough Set Data Analysis, IEEE GrC 2012, 2012 年 8 月 11 日～13 日，杭州（中国）．

[6] 工藤康生，岡田隆生，村井哲也，ラフ集合におけるヒューリスティックな縮約計算の並列化の試み，第 7 回日本感性工学会春季大会，2012 年 3 月 3 日～4 日，高松(日本)．

[7] Yasuo Kudo, Ken Kaneiwa, and Tetsuya Murai, An Attempt of Reconstruction of Object-Oriented Rough Set Models, IEEE GrC2011, 2011 年 11 月 8 日～10 日，高雄(台湾)．

６．研究組織
(1)研究代表者
工藤　康生　（KUDO YASUO）
室蘭工業大学・工学研究科・准教授
研究者番号：90360966