様式 C-19

科学研究費補助金研究成果報告書

平成２２年５月２１日現在

研究種目：基盤研究（B）
研究期間：2007～2009
課題番号：１９３０００７４
研究課題名（和文）ラフセット・クラスタリング技法の確立-ラフ集合論の新たな局面の開拓

研究課題名（英文）Establishing the method of rough set clustering

研究代表者
宮本　定明（ MIYAMOTO SADAAKI ）
筑波大学・大学院システム情報工学研究科・教授
研究者番号：60143179

研究成果の概要（和文）：
ラフ集合は集合の近似を行う理論であり，典型的な応用例は情報表と決定表に関わる近似である．本研究では，従来教師付き分類を論じてきたラフ集合論に加えて，ラフ集合に関わるクラスタリングの理論を，理論・アルゴリズム・応用の 3 つの側面から総合的に研究し，ラフセット・クラスタリングの理論の可能性を追求した．従来は単発的に実施されてきたラフ集合のクラスタリングに対して，現在存在する主要なクラスタリング技法である階層的技法，K-平均法に関連する技法，カーネルクラスタリング技法などについて，それらとラフ集合およびその一般化との関連を理論的に明らかにし，そのことから導出される様々なアルゴリズムを開発し，医療情報関連データを含むいくつかの応用を行った．このことで，ラフ集合に関わるクラスタリング技法の全貌がほぼ明らかになった．

研究成果の概要（英文）：
Traditional rough set theory is concerned with supervised classification of decision tables, with the exception of a few proposals of unsupervised classification, in other words, clustering related to rough sets. However, there are many clustering techniques that had not been applied to rough sets before this study. We have studied theoretical background of clustering for rough sets, and developed a number of new algorithms including both hierarchical and non-hierarchical clustering. We moreover applied the developed methods to a number of data sets including those in medical information. We thus have shown a broad overview of clustering techniques of rough sets and related applications.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	5,500,000	1,650,000	7,150,000
2008 年度	4,300,000	1,290,000	5,590,000
2009 年度	4,500,000	1,350,000	5,850,000
年度			
年度			
総　計	14,300,000	4,290,000	18,590,000

研究分野：ソフトコンピューティング基礎論，クラスタリングアルゴリズム
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：ラフ集合，階層的クラスタリング，一般化ラフ集合，情報表，K-平均法

１．研究開始当初の背景

ラフ集合は元来教師付き分類に関連した概念であるが，クラスタリング技法の必要性も認識され，いくつかの研究が単発的に行われていた．それらは，Lingras らによるラフ K-平均法（rough K-means）であり，平野らによるラフ・クラスタリングであった．ところが前者は，ユークリッド空間内のデータ集合に対する技法であるので，ラフ集合が扱う対象である情報表や決定表をクラスタリングするには適切ではない．また，後者のラフ・クラスタリング技法は独創的ではあるが，複雑な処理を必要とするので，大量のデータを扱うには向いていない．また，根本的な疑問点として既存の標準的なクラスタリング技法が，ラフ集合論における諸概念とどのように関係しているのか，また，適用できるのか，が明らかではなかった．

２．研究の目的

本研究では，このような背景に鑑み，ラフ集合に関わるクラスタリングの様々な可能性を究めることを目的として，標準的なクラスタリング技法がラフ集合が対象とするデータにどのように適用されるかを明らかにし，かつそれに関わる理論的問題を解明し，いくつかの応用を行うことを目的とした．その際，既存の方法が直接的に適用できるのではなく，新規の技法が開発されなければならない点に注意する．このことによって，ラフ集合に関わるクラスタリング技法の全体像を明らかにすることが本研究の目的である．

３．研究の方法

本研究では，クラスタリングに関する理論的考察とそれに基づくアルゴリズムの開発，情報表と決定表のクラスタリング，医療情報データのクラスタリングを含むいくつかの応用，の３つの面から研究を進める．

(1) クラスタリングに関する理論的考察とそれに基づくアルゴリズムの開発

代表者と分担者村井を中心とする研究班では，階層的あるいは非階層的クラスタリングによって情報表をクラスタリングする際の基本的問題に焦点をあて，理論的問題の解決を試みるとともに，理論的考察から導出される新たなクラスタリングアルゴリズムを開発する．

(2) 情報表と決定表のクラスタリング

分担者乾口を中心とする研究班は，情報表と決定表のクラスタリングに焦点を当てて，アルゴリズムを開発し，その効果を調べる．

(3) 医療情報データ等への応用

分担者津本を中心とする班は，医療情報データへのクラスタリングの応用を中心に，考察を進める．また，平野・津本のラフ・クラスタリング技法を発展させる．

これらの研究を実施するにあたって．情報交換を密にすると同時に，各自の分担領域を超えて互いに協力し，新規性・有用性のある技法の研究開発に努める．

４．研究成果

以下の研究成果を得る以前に，当該研究の最初の年度において，ラフ集合に関する国際学会 JRS2007 における基調講演を行い，ラフ集合のクラスタリングに関する現状を概観すると同時に，既存の方法の適用可能性を示し，あわせてその問題点を述べた（論文⑮）．

(1) ラフセット・クラスタリングの基礎理論

以下の３つの側面から，ラフセット・クラスタリングの基礎となる理論について考察した．

① 順序集合にもとづく一般化階層的技法

従来の階層的クラスタリング技法における併合レベルを一般化することによって，束をはじめとする順序集合にもとづく一般化クラスタリングの技法を提案し，情報表との関係を考察した（論文⑮他）．また，分担者村井を中心に，この数学モデルの基礎となる位相的構造の考察を行った（学会発表④）．さらに，階層的クラスタリング技法を一般化する際に生じる理論的問題として，細分の理論，たとえばある技法によるクラスターが別の技法によるクラスターの細分になるかどうか，を考察し，最短距離法と他の方法の間では満足すべき理論的性質が成り立つが，他の方法の間ではこれに対応する性質は成り立たないことを反例により示した（論文④）．

② 決定表と分割表

決定表から得られる分割表は，クラスタリングのための類似度を定義するために用いられる．そこで，分担者津本らは，分割表における統計的独立性を線形代数の観点から考察し，それらの関係を明らかにするいくつかの理論的性質を導出した（論文⑨）．

③ 情報表と決定表における類似性

分担者乾口らが中心となり，ラフ集合論における基本的要素である情報表や決定表を複数与えたときのそれらの間の類似性の測

度を考察し，多様な情報表および決定表をグループ化する際の基礎となる理論的性質を導出した．また，この考察から定義される類似性の測度をクラスタリングアルゴリズムに用いるための方法論についての提案を行った（論文⑦）．

④ ファジィ理論とラフセット・クラスタリング

ファジィ理論が，ラフ集合のクラスタリングのどのように役立つかについて，次の２つの面から考察した．一つは，ファジィラフ集合が，階層的方法における最短距離法と密接に関係していることであり，いま一つはファジィ c－平均法が，これまで研究されてきたラフ c－平均法の一般化として位置づけられことである．これらの性質を理論的観点から明らかにした（論文③）．

(2) ラフセット・クラスタリング技法の開発

① 一般化ラフ近似にもとづくクラスタリング技法

ラフ近似を一般化する研究がこれまで多く行われている．ラフセット・クラスタリングにおいては，ファジィラフ近似や近傍システムなどの一般化ラフ集合の概念が適用できる．

本研究では，これらのモデルのもとで適用すべきクラスタリング技法として，カーネル関数法を考察した．ファジィラフ近似や一次元のファジィ近傍システムについて，正定値カーネルが得られるための条件を示し（論文⑭，学会発表⑥），階層的および K-平均型のクラスタリングアルゴリズムを導出した（論文⑤，学会発表⑤）．また，個々のデータに近傍を伴うデータ集合のクラスタリング技法を，制約付き最適化技法にもとづいて検討し，通常のクラスタリング技法と回帰モデルを含むクラスタリングアルゴリズムを開発した（学会発表①）．

② クラスター数の問題の考察

上記の近傍概念に基づくクラスタリングアルゴリズムを利用する際や，後に述べる決定表の K-平均型クラスタリングにおいては，仮定するクラスター数が問題となる．本研究では，クラスター数を仮定しない逐次クラスター抽出アルゴリズムを提案した（論文⑩）．その一方で，クラスター妥当性検証のための基準が，近傍モデルなどで利用するカーネル関数を使用した場合にも良好なパフォーマンスを示すかどうかを，カーネルを利用した妥当性基準を提案し，かつ多くの数値例によるシミュレーションによって検証した．この際，従来提案されてきたいくつかの異なる妥当性基準をカーネル化した場合としない場合とで比較した．結論として，カーネル関数を利用した場合の妥当性基準は，利用しない場合とあまり変わらない性能を示し，使用に耐えること，これまで提案されてきた妥当性基準のうち，たとえば Fukuyama-Sugeno 基準の性能は，カーネル関数を用いた場合は特に劣ること，クラスター共分散の行列式を利用した基準はカーネルクラスタリングでは使用できないこと，利用するクラスタリング技法としては，Dunn，Bezdek によるファジィ c-平均法がハード K-平均法よりも優れていると判定できることなどがわかった（論文⑥）．

③ラフ・クラスタリングに関する研究

分担者平野・津本らによるラフ・クラスタリング技法を発展させ，非ユークリッド的性質をもつ非類似度や非対称類似度に対する技法を開発した（論文⑬）．また，決定表にもとづいた分割表を利用した類似度の代表例として，Jaccard 係数を用いた階層的アルゴリズムを提案した（論文①）．

④ 2 重メンバーシップクラスタリング技法の提案

ラフ近似では，上近似と下近似の 2 種類が考察され，Lingras らによる Rough K-平均法でも，クラスターの上近似と下近似の概念を用いる．本研究では，これらを一般化した 2 重メンバーシップクラスタリング技法を開発した．2 重メンバーシップは，Dunn,Bezdek の方法を変形し，異なる 2 種類のファジィ化パラメータを同時に用いて得られる．この際，2 つのクラスタリングアルゴリズムを動かすのではなく，2 つのパラメータをもつ単一のアルゴリズムが動作する．このアルゴリズムを提案し，その性質を明らかにするため，変分法を利用した理論的解析を行った（論文⑪）．

(3) 決定表のクラスタリング

①単一の情報表に対するクラスタリング

単一の情報表に対して，K-平均法を適用する場合には，K-モード法を用いるのが最も直接的な方法であることを示した．また，その際の問題点と一般化の方法を明らかにした（論文⑮他）．

②複数の情報表と決定表に対するクラスタリング

モデルベース階層的クラスタリングの概念は，研究代表者らによって，既に提案されていたが，ここでは，複数の決定表をグループ化するための類似度をいくつか提案し，クラスタリング技法を開発した．すなわち，階層的クラスタリングの新たなアルゴリズムを提案し（論文⑫）．さらに，K-平均型のアルゴリズムを開発した（学会発表②）．

(4) 医療情報への応用
①ラフ・クラスタリングの利用
分担者平野・津本らによるラフ・クラスタリング技法を医療データに関する情報表に適用し，ラフ近似などの教師付き分類などと合せて医療リスクの問題として総合的に考察し，その中でのクラスタリングの位置づけを明らかにした（論文②）.

②近傍システムの医療インシデントレポート解析への応用
ファジィ近傍概念にもとづいて開発したカーネルクラスタリング技法とカーネル主成分分析技法を医療インシデントレポートデータに適用し，クラスターと主成分を抽出した（論文⑤）．この技法では，キーワードのクラスターが生じるが，現時点で扱っているキーワードは未だ数が限られている．しかしながら，技法としては，基本的に完成の段階に達してしており，今後様々なデータに適用することができる.

(5) その他の応用
①画像情報検索への応用
ラフ集合における上近似・下近似およびマルチラフ集合モデルにもとづくクラスタリングの技法を画像情報の索引付けと検索に応用する方法を提案した（論文⑯）.

②データプライバシーへの応用
データプライバシーは，現在注目を集めている分野である．情報表のクラスタリング技法をデータプライバシーに応用する方法を提案した（論文⑧）.

③グループ意思決定への応用
階層的クラスタリングによって，決定表の類似性を分析し，グループ意思決定に応用する方法を示した（論文⑫，学会発表③）.

(6) まとめと今後の展望
はじめに述べたように，本研究がはじまる前には，ラフ集合のクラスタリング技法は，ユークリッド空間に対するラフ K-平均の技法と連携研究者平野らによるラフ・クラスタリング技法しか研究されていなかった．本研究では，一般化階層的クラスタリング技法，一般化ラフ集合に対するカーネルクラスタリング技法，2 重メンバーシップクラスタリング技法，ハード c-平均法とファジィ c-平均法の情報表への応用，さらに複数の情報表の階層的・非階層的クラスタリング技法を新たに開発した．またこれらに関連する理論的諸問題を解決し，クラスター妥当性基準のような経験的問題はシミュレーションにより考察した．ラフ・クラスタリングに関しては，多様な形式をもつデータに応用できるように，技法を一般化した.
応用面については，当初の目的である医療関連データへの応用を行っただけでなく，画像検索へのラフ集合モデルの適用や情報表のデータプライバシーへの応用，グループ意思決定への応用について考察した.
これらの成果にみるように，ラフセット・クラスタリングに関するアプローチと技法を網羅的かつ総合的に研究開発した．そのことによって，当初の研究目的は十分に果たされた．のみならず，理論面・応用面の両方で予想を超えた成果が挙げられた．また，ラフ集合の分野では，他の研究グループでは，本質的に新しいクラスタリング技法は提案されていないため，質・量ともに，他の研究グループを圧倒する研究成果が得られている.
また，当該研究の最終年度頃になって，グラニュラークラスタリング（粒状クラスタリング）という概念がラフ集合論分野とは異なる統計・人工知能分野の研究者の間で提案されつつある．本研究による成果は，ラフ集合分野のみならず，今後の発展が予想されるグラニュラークラスタリングの研究領域においても，大きなインパクトを示すものである.

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 28 件）

① S. Hirano, S. Tsumoto, Representation of Granularity for Non-Euclidian Relational Data by Jaccard Coefficients and Binary Classifications, Proc. of RSCTC 2010, to appear, 査読有.
② S. Tsumoto, S. Hirano, Risk Mining in Medicine: Application of Data Mining to Medical Risk Management, Fundamenta Informaticae, Vol.98, pp. 107-121 (2010)，査読有.
③ 宮本定明，ファジィクラスタリングの有用性について，知能と情報（日本知能情報ファジィ学会誌），Vol.21, No.6, pp.1008-1017（2009），査読有.
④ S. Miyamoto, Refinement Properties in Agglomerative Hierarchical Clustering, LNAI5861, pp. 259-267 (2009)，査読有.
⑤ 河崎佑一，宮本定明，ファジィ近傍に基づく内積空間とカーネル法によるテキストデータ解析，知能と情報（日本知能情報ファジィ学会誌），Vol.21, No.4, pp.461-469（2009），査読有.
⑥ W. Hashimoto, T. Nakamura, S. Miyamoto, Comparison and Evaluation of Different Cluster Validity Measures Including

Their Kernelization, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.13, No.3, pp.204-209 (2009), 査読有.

⑦ M. Inuiguchi, Rough Set Approach to Rule Induction from Imprecise Decision Tables, Fuzzy Logic and Applications, 8th International Workshop, WILF 2009, LNAI 5571, pp.68-76 (2009), 査読有.

⑧ V. Torra, Y. Endo, S. Miyamoto, On the comparison of some fuzzy clustering methods for privacy preserving data mining, Kybernetika, Vol.45, pp. 548-560 (2009), 査読有.

⑨ S. Tsumoto, S. Hirano, Statistical Independence and Determinants in a Contingency Table - Interpretation of Pearson Residuals based on Linear Algebra, Fundamenta Informaticae, Vol.90, pp. 251-267 (2009), 査読有.

⑩ S. Miyamoto, Y. Kuroda, K. Arai, Algorithms for Sequential Extraction of Clusters by Possibilistic Method and Comparison with Mountain Clustering, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.12, No.5, pp.448-453 (2008), 査読有.

⑪ S. Miyamoto, Formulation of Fuzzy c-Means Clustering Using Calculus of Variations and Twofold Membership Clusters, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.12, No.5, pp.454-460 (2008), 査読有.

⑫ R. Enomoto, M. Inuiguchi, Group Forming of Decision Tables Based on Rational Choice Theory, Proc. SCIS & ISIS 2008, pp.1002-1007 (2008), 査読有.

⑬ S. Hirano, S. Tsumoto, Hierarchical Clustering of Non-Euclidean Relational Data Using Indiscernibility-Level. Third International Conference Rough Sets and Knowledge Technology (RSKT 2008), Chengdu, China, May 17-19, pp. 332-339 (2008), 査読有.

⑭ S. Miyamoto, Y. Kawasaki, Kernel Functions Based on Fuzzy Neighborhoods and Agglomerative Clustering, V. Torra, Y. Narukawa, and Y. Yoshida (Eds.): MDAI 2007, LNAI 4617, pp.249-260, (2007), 査読有.

⑮ S. Miyamoto, Data Clustering Algorithms for Information Systems, Proc. of RSFDGrC2007, LNAI 4482, Springer, pp.13-24 (2007), 査読有.

⑯ T. Murai, S. Miyamoto, Y. Kudo, A Logical Representation of Images by Means of Multi-rough Sets for Kansei Image Retrieval, LNCS 4481, pp.244-251 (2007), 査読有.

〔学会発表〕(計 45 件)

① K. Kurihara, Y. Endo, Y. Hamasuna, S. Miyamoto, Fuzzy c-Regression Model for Data with Tolerance, MDAI2009, December 1, 2009, Awaji Island, Japan.

② 乾口雅弘, 榎本隆太, 決定表の非階層的クラスタリング, 第52回自動制御連合講演会, 2009年11月22日, 大阪大学豊中キャンパス.

③ 榎本隆太, 乾口雅弘, 合理的選択理論による決定表のグループ形成解析, 24th Fuzzy System Symposium (Osaka, September 3, 2008).

④ 村井哲也, 三浦博史, 生方誠希, 工藤康生, 宮本が提案した順序集合に基づくクラスタリングに関する位相の観点からの一考察. 24th Fuzzy System Symposium (Osaka, September 3, 2008).

⑤ S. Miyamoto, Recent Studies on Algorithms for Fuzzy Clustering, Proc. of 2008 IEEE International Conference on Granular Computing (GrC 2008), pp. 59-60, Aug. 26, 2008, Hangzhou, China.

⑥ S. Miyamoto, Y. Kawasaki, Kernel Space for Text Analysis Based on Fuzzy Neighborhoods, WCCI 2008 Proceedings, 2008 IEEE World Congress on Computational Intelligence, June 4, 2008, Hong Kong, pp.738-743, (2008).

６．研究組織

(1)研究代表者

宮本　定明 (MIYAMOTO SADAAKI)

筑波大学・大学院システム情報工学研究科・教授

研究者番号：　60143179

(2)研究分担者

津本　周作 (TSUMOTO SHUSAKU)

島根大学・医学部・教授

研究者番号：10251555

乾口　雅弘 (INUIGUCHI MASAHIRO)

大阪大学・大学院基礎工学研究科・教授

研究者番号：60193570

村井　哲也 (MURAI TETSUYA)

北海道大学・大学院情報科学研究科・准教授

研究者番号：90201805

(3)連携研究者

平野　章二（HIRANO SHOJI）
島根大学・医学部・准教授
研究者番号：60333506