様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　　５　日現在

機関番号：　１４４０１
研究種目：　若手研究(B)
研究期間：　2011～2014
課題番号：　２３７００２６５

研究課題名（和文）ラフ集合理論の属性縮約に基づいた非類似度によるクラスター分析

研究課題名（英文）Cluster Analysis Using Dissimilarity Based on Attribute Reduction of Rough Set Theory

研究代表者

楠木　祥文（KUSUNOKI, Yoshifumi）

大阪大学・工学（系）研究科（研究院）・助教

研究者番号：３０５８８３２２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：ラフ集合理論の属性縮約は，対象集合の識別可能性/不能性に基づき，データから冗長な属性を取り除く手法である．本課題は，名義的データに対して，識別可能性に基づく二つの類似度/非類似度を提案した．一つは，クラスター間の非類似度であり，その値は二つのクラスターを識別する属性部分集合の数で定義される．もう一つは，識別可能性を反映した特徴空間に対するカーネル関数であり，その特徴空間は与えられた情報と整合する属性部分集合族である．これらの類似度/非類似度をクラスタリングと決定ルール抽出に適用した結果，分類性能と表現の簡潔さを両立させるようなクラスターや決定ルール群が得られることが実験的に明らかになった．

研究成果の概要（英文）：Attribute reduction of rough set theory is a methodology to remove irrelevant attributes from a data set, which is based on discernibility/indiscernibility of object sets. In this research, we have proposed two kinds of similarity/dissimilarity of objects based on the discernibility for nominal data sets. Moreover, we have proposed data analysis methods using them. One is dissimilarities for clusters, which are defined by the number of attribute subsets discerning two clusters. The other is kernel functions reflecting discernibility, whose feature spaces are discerning attribute subsets. We have applied those similarities and dissimilarities to clustering and decision rule induction tasks. It is shown by numerical experiments that we can obtain clusters and decision rules balancing classification accuracy and simplicity using proposed approaches.

研究分野：　データマイニング

キーワード：　データマイニング　機械学習　カーネル法　論理関数　ラフ集合

1. 研究開始当初の背景

データ分析における主要はタスクとして，対象とする物事の分類があげられる．一般的なデータ分析では，与えられた対象（または事例，観測点など）は，複数の属性によって記述され，属性に基づき定義された類似度/非類似度によって，対象は分類される．属性は名義（非数値，カテゴリカル）属性と数値属性に大きく分けられるが，本研究課題では名義属性のみで構成されたデータ（名義的データ）に着目する．名義的データの分析では，対象のクラスラベルを表現する識別関数を求めることや，対象のクラスターを発見することだけではなく，得られた識別関数やクラスターが簡潔な論理表現を持つことも重要である．

ラフ集合理論は対象の識別可能/不能性に基づき，集合の矛盾，曖昧さを取り扱うモデルである．その矛盾に着目することで，属性の重要性の概念が提案され，属性縮約に利用されている．本研究課題では，ラフ集合の属性縮約の考えに基づいた類似度/非類似度，および，それを用いたデータ分析手法を提案する．本アプローチにより，簡潔な論理表現を持つ識別関数やクラスターが得られると考えられる．

名義的データに対する分析は，医療診断，感性工学，意志決定分析，生物分類学，人文・社会科学など，多岐に渡る分野で見られ，本アプローチの有効性が示された場合，その影響は大きい．

2. 研究の目的

本課題は名義的な値を持つデータに対して，識別可能性/不能性に基づいた新たな分析手法の開発を目的としている．属性集合における対象の識別不能関係とは，その属性集合のみを用いた場合，二つの対象が全く同じ表現となり，識別できないことをいう．識別可能性または識別不能性は，重要な属性を求める属性縮約や，データに内在する規則を発見するルール抽出などのラフ集合理論に基づくデータ分析手法において重要な役割を果たしている．本課題では，識別可能性/不能性を反映した対象の類似度/非類似度を提案し，データ分析手法に応用する．

3. 研究の方法

本課題では，識別可能性/不能性に基づく類似度/非類似度を二つ提案し，その性質を調査する．一つは，(1) 識別可能性を用いたクラスター間の非類似度であり，その値は二つのクラスターを識別する属性部分集合の数で定義される．もう一つは，(2) 識別可能性を反映した特徴空間に対するカーネル関数（内積）であり，その特徴空間は与えられた情報と整合する属性部分集合の族で定義される．

これらの類似度/非類似度を用いたクラスタリングとルール抽出について研究を行う．クラスタリングとは対象の集合を，属性に基づく類似度・非類似度によって，外的基準なしに分類する手法である．ルール抽出とは，複数のクラスに分割されたデータから，各クラスを推論する if-then ルール（決定ルール）を求める手法である．決定ルールの条件部は属性値に関する制約を表しており，データに内在する規則とみなすことができる．一般に，クラスタリングは教師なし学習，ルール抽出は教師あり学習と呼ばれる．

4. 研究成果

(1) クラスター間の識別不能性に基づいた非類似度とそれを用いたデータ分析手法の提案

一つ目の成果として，識別可能性に基づいたクラスター間の類似度を提案し，それをクラスタリングに応用した．従来のクラスタリングでは，まず対象間の非類似度が属性値の不一致度などによって定義され，平均，最大値，最小値などによってそれをクラスター間の非類似度に拡張する．名義属性を持つ対象集合のクラスタリングでは，類似した対象で構成されたクラスターを求めるだけでなく，得られたクラスターが，例えば「(v1 = 1 かつ v2 = 1) あるいは (v3 = 2)」のような，簡潔な論理的表現（パターン）を持つことが重要である．そのためには，各属性についての対象間の一致・不一致だけではなく，属性部分集合におけるクラスター間の識別可能/不能性も考慮する必要がある．ここで，クラスター間の識別可能性とは，ある属性部分集合において，異なるクラスターに含まれる任意の二つの対象を識別できることを指す．提案する非類似度はクラスター間を識別できる属性部分集合の数によって定義される．

しかし，ある属性部分集合でクラスターが識別可能であれば，それを包含する属性部分集合でも識別可能であるため，クラスターが小さな部分集合で識別されれば，異なるクラスターに属する対象間の不一致度がどれだけ大きくても，非類似度は小さくなってしまう．そこで，対象間の不一致度を考慮することで，識別可能性に基づく非類似度を改良した．

また，提案した非類似度の弱点として，計算量の大きさがあるが，この非類似度を階層的クラスタリングに適用した場合について，計算を削減する方法を提案した．

提案した非類似度について，その性質を考察

するとともに，それを階層的クラスタリングに適用し，実験的にその有用性を評価した．従来法と比較して，提案法では，いくつかのデータにおいて，群内距離と簡潔さを両立させるクラスターが得られることが分かった．

二つ目の成果として，上述の非類似度を応用したルール抽出を提案した．従来のルール抽出法は逐次被覆法に基づいている．逐次被覆法では，まず，目標とするクラスに含まれる対象を決定ルールが被覆すべき正対象と定め，それ以外の対象を負対象とする．そして，その目標クラスを推論する決定ルールの条件部を逐次生成する．一つの条件部を求めるとき，ほとんどのルール抽出アルゴリズムでは，一般/特殊アプローチを用いている．これは，条件部が空の決定ルールからはじめて，負対象を（ほとんど）被覆しなくなるまで条件部に新たな条件を追加していく方法である．本課題では，識別可能性に基づく非類似度を決定ルール抽出に適用することで，特殊/一般アプローチによるルール抽出法を提案した．つまり，ある対象に対応する条件部を持つ決定ルールからはじめて，正対象を被覆する範囲で，条件を削除していくことで，決定ルールを得る．提案手法では，クラスターとその補集合の非類似度をクラスターの評価値として用い，正対象のクラスターを生成する．それに含まれる対象に共通する条件属性値を用いることで，決定ルールを得る．

数値実験により，提案法は，従来のルール抽出法の LEM2 よりも，簡潔な決定ルール群を抽出できることが分かった．しかし，LEM2 と比べ分類精度に関して劣る傾向が見られる．この原因として，提案法は過学習を起こしていると推測し，それを回避する方法を提案した．この改良により，提案法で得られる決定ルール群の分類精度を向上させることができた．

(2) 論理関数とカーネル法を用いたデータ分析手法の提案

三つ目の成果として，識別可能性を反映したカーネル関数を提案し，それを制約付きクラスタリングに用いた．制約付きクラスタリングとは，同じクラスターに含まれるべき対象ペア（must-link）や，異なるクラスターに分類されるべき対象ペア（cannot-link）が与えられた下でのクラスタリングである．ここでは cannot-link のみを考慮する．上の(1)で述べたとおり，名義的データに対するクラスタリングでは，類似した対象によるクラスターを発見するだけではなく，そのクラスターを記述できる論理表現（パターン）についても考慮する必要がある．cannot-link が与えられた場合，それに整合しないパターンを考慮したクラスタリング手法が望まれる．そのために，まず，各対象 x をパターン空間上の下側論理関数 h に写す．パターン p が x に満たされるとき h(p)=1 とする．つまり，h は x に満たされるパターンの集合を表している．対象の各ペアについて，パターン空間上の内積を計算すれば，カーネル法によって，対象集合に対するクラスタリングを実行できる．さらに，下側論理関数の定義域から，cannot-link の二つの対象に同時に満たされるパターンを排除することで，cannot-link の情報をクラスタリングに反映する．提案するカーネル関数（内積）は制限された下側論理関数（RDF: Restricted Downward Function）カーネルと呼ばれる．

人工データなどを用いた数値実験では，適切なクラスターを得ることができた．また，RDF カーネルに対応するパターンの空間対する主成分分析による 2 次元表現によって，cannot-link を反映した対象の配置が得られていることを確認した．

四つ目の成果として，上述の RDF カーネルを用いた教師付き学習を提案した．教師付き学習では，クラス情報が対象集合の分割として与えられるが，異なるクラスの対象のペアを cannot-link とし，上述のようにカーネル関数を定義すると，クラス情報を反映した特徴空間を得ることができる．つまり，特徴空間では，異なるクラスの対象を同時に満たされるパターンが排除される．このカーネル関数を用いて線形識別関数を構成する．このとき，識別関数はクラス情報と整合する全パターンの重み付き和で表現でき，決定ルールによる識別器とみなすことができる．つまり，決定ルール抽出を暗に行っていることになる．

RDF カーネルと Support Vector Machine を用いて線形識別器を構成した．RDF カーネルを評価するために，不整合なパターンを排除しない冪集合カーネルによる識別器と比較した．サンプルから真の論理関数を推定する問題では，冪集合カーネルと比較して，RDF カーネルを用いることで，真の論理関数により近い識別器が得られることが分かった．

5. 主な発表論文等

〔学会発表〕（計 18 件）

① 木邑 明倫，制限された下側論理関数カーネルを用いた分類器の構築，第 59 回システム制御情報学会研究発表講演会，2015 年 5 月 20 日～22 日，中央電気倶楽部（大阪府・大阪市）．

② 楠木 祥文，制限された下側論理関数カーネルとサポートベクトルマシンを用いた部分定義論理関数の拡張，日本オペレーションズ・リサーチ学会 2015 年春季研究発表会，2015 年 3 月 26～27 日，東京理

科大学神楽坂キャンパス（東京都・新宿区）.

③ 楠木 祥文,制限された下側論理関数カーネルの特徴空間について，計測自動制御学会関西支部・システム制御情報学会若手研究発表会，2015 年 1 月 16 日，大阪大学吹田キャンパス（大阪府・吹田市）.

④ 楠木 祥文,制限された下側論理関数カーネルを用いた制約付きクラスタリング，第 57 回自動制御連合講演会，2014 年 11 月 10～12 日，ホテル天坊（群馬県・渋川市）.

⑤ Y. Kusunoki, Boolean Kernels and Clustering with Pairwise Constraints, 2014 IEEE International Conference on Granular Computing, 22-24 Oct. 2014, 登別グランドホテル（北海道・登別市）, pp. 141-146, DOI: 10.1109/GRC.2014.6982823

⑥ Y. Kusunoki, Boolean Kernel Functions for Nominal Data Analysis, 17th Czech-Japan Seminar on Data Analysis & Decision Making, 16-19 Sep. 2014, 北九州国際会議場（福井県・北九州市）.

⑦ 楠木 祥文,クラスターの識別可能性を反映した属性部分集合カーネル,第 30 回ファジィシステムシンポジウム，2014 年 9 月 1～3 日，高知城ホール（高知県・高知市）.

⑧ 田中 大樹,識別可能クラスタリングに基づく決定ルール抽出における過学習の回避,第 30 回ファジィシステムシンポジウム，2014 年 9 月 1～3 日，高知城ホール（高知県・高知市）.

⑨ 楠木 祥文,名義的データに対するクラスタリングのためのカーネル関数,第 24 回ソフトサイエンス・ワークショップ,2014 年 3 月 8～9 日，久留米ホテルエスプリ（福岡県・久留米市）.

⑩ Y. Kusunoki, Specific-to-general Approach for Rule Induction Using Discerniblity Based Dissimilarity, 2013 IEEE International Conference on Granular Computing, 13-15 Dec. 2013, Beijing (China), pp. 178-181, DOI: 10.1109/GrC. 2013.6740403.

⑪ Y. Kusunoki, Hierarchical Clustering Using Proximity Measures Based on Discernibility of Clusters, 16th Czech-Japan Seminar on Data Analysis and Decision Making, 19-22 Sep. 2013, Marianske Lazne (Czech Republic).

⑫ 楠木 祥文,識別可能性に基づくクラスタリングを用いた決定ルール抽出について，第 29 回ファジィシステムシンポジウム，2013 年 9 月 9～11 日，大阪国際大学（大阪府・枚方市）.

⑬ 楠木 祥文,完全不一致を考慮した識別可能性に基づく非類似度,第 28 回ファジィシステムシンポジウム,2012 年 9 月 12～14 日，名古屋工業大学（愛知県名・古屋市）.

⑭ Y. Kusunoki, Dissimilarity Based on Cluster Discernibility on Attribute Subsets, 23th European Conference on Operational Research, 8-11 Jul. 2012, Vilnius (Lithuania).

⑮ 梅村 一紀,識別可能性に基づく非類似度に対する効率的な比較手法,第 56 回システム制御情報学会研究発表講演会，2012 年 5 月 21 日～23 日，京都テルサ（京都府・京都市）.

⑯ Y. Kusunoki, Agglomerative Hierarchical Clustering with Dissimilarity Using Discernibility on Attribute Subsets for Nominal Data Sets, 2011 IEEE International Conference on Granular Computing, 8-10 Nov. 2011, Kaohsiung (Taiwan), pp. 357-362, DOI: 10.1109/GRC.2011.6122622.

⑰ Y. Kusunoki, Dissimilarities Using Discernibilities of Objects on Attribute Subsets for Clustering, 14th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty, 18-21 Sep. 2011, Hejnice (Czech Republic).

⑱ 楠木 祥文,属性部分集合族におけるクラスターの識別可能性を用いた非類似度，第 27 回ファジィシステムシンポジウム，2011 年 9 月 12～14 日，福井大学（福井県・福井市）.

6. 研究組織

(1) 研究代表者

楠木　祥文（KUSUNOKI, Yoshifumi）
大阪大学・大学院工学研究科・助教
研究者番号：３０５８８３２２