情報表の縮約を用いたクラスタリング手法について

On a Clustering Method Using Reducts of Information Tables

中西 基　工藤 康生　村井 哲也
Hajime Nakanishi　Yasuo Kudo　Tetsuya Murai
室蘭工業大学　北海道大学
Muroran Institute of Technology　Hokkaido University

Abstract: In this paper, we propose a clustering method using reducts of information tables in rough set theory. In the proposed method, we perform a clustering by choosing attributes with low ratio of appearance in reducts. Experimental results indicate that the proposed method enables us to classify the data without introducing similarity or non-similarity measures and deciding the number of decision classes.
Keywords： rough set, clustering, reduct

1　はじめに

クラスタリングは近年盛んに研究されており，様々な手法が提案されている (詳細は文献 [3] 参照)．一般的なクラスタリング手法では，データ間の類似性あるいは非類似性を表す測度を用いてデータの分類を行う．また，k-means 法などの非階層的クラスタ分析法では，あらかじめ分割するクラスタの個数を決める必要がある．本研究では，情報表として与えられたカテゴリカルなデータに対して，ラフ集合理論 [5] における情報表の縮約を用いることで，類似性あるいは非類似性を表す測度を導入することなく，かつクラスタの個数をあらかじめ決めずにデータをクラスタリングする手法を提案する．

2　情報表と縮約

文献 [4] に基づいて，情報表と縮約について概説する．ラフ集合の手法を用いる場合，対象に関するデータは複数の属性とそれらの値で与えられることが多い．多くの対象に対する属性値データを示した表を**情報表**と呼ぶ．情報表は，(U, AT, V, ρ) の 4 対で定義される．U は情報表に現れる対象全体の集合，AT は属性の集合，V は属性 a のとる値の集合 V_a を用いて，$V = \bigcup_{a \in AT} V_a$ と定められ，$\rho : U \times AT \to V$ は対象 u と属性 a に対して属性値 $\rho(u, a) \in V$ を割り当てる関数である．属性の任意の部分集合 $A \subseteq AT$ が与えられると，情報表に基づき，次の関係 R_A を定めることができる．

$$R_A = \{(x, y) \mid \rho(x, a) = \rho(y, a), \forall a \in A\}.$$

$(x, y) \in R_A$ であるとき，対象 x と y が属性の部分集合 A により識別できないことを表しているので，R_A は**識別不能関係**という．R_A は，反射性，対象性，推移性を満たすので，同値関係である．

情報表で与えられた全ての属性の集合 AT と同等に対象を識別できるために必要な最小の属性の部分集合を**縮約**と呼ぶ．縮約は次のように定義される．

$$R_A = R_{AT} \text{かつ } \nexists a \in A; R_{A-\{a\}} = R_A.$$

上の条件を満たす $A \subseteq AT$ が縮約である．

3　情報表の縮約を用いたクラスタリング

本研究では，情報表として与えられたカテゴリカルなデータに対するクラスタリングとして，情報表の縮約を用いたクラスタリング手法を提案する．縮約を用いて良いクラスタリングを行うためには，データの特徴を強く反映した属性を抽出することが望ましい．そこで，データの特徴を強く反映した属性と属性が縮約に含まれている割合 (出現率) との関係を調べるため，予備実験を行う．

予備実験に用いるデータはサンプル数 30，属性数 10，属性値は全て 0 か 1 の情報表である．このデータの 1〜3 番目の属性は規則性のある属性値，それ以外はランダムに生成した規則性のない属性値を持つ．また各属性値に対して 0 から 1 までの乱数を発生させ，乱数が事前に設定したノイズ発生率以下の属性値は，その値を反転させている．予備実験ではノイズ発生率を 0 から 0.01 ずつ上昇させ，各ノイズ発生率において，生成した情報表のすべての縮約を求め，規則性を反映した属性とそうでない属性の縮約における出現率を求める試行を 100 回行い，その平均値を求めた．実験結果を図 1 に示す．

濃い線が規則性ありの属性の平均値で，薄い線が規則性なしの属性の平均値である．図 1 より，ノイズ発生率

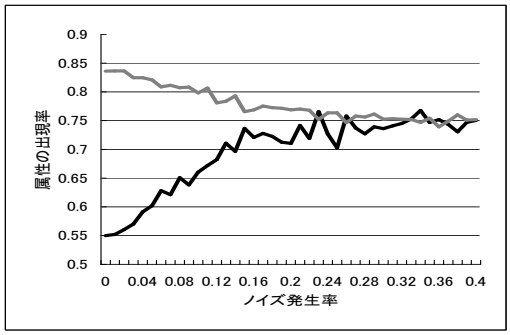

図 1: 予備実験の結果

が小さい場合，規則性のある属性は規則性のない属性よりも出現率が低くなる傾向が見られた．よって，何らかの特徴を持ったデータの場合，出現率の低い属性を選べば，簡潔で特徴のあるクラスタを作ることができると思われる．

以上の実験結果を考慮して，本研究では以下のアルゴリズムで情報表として与えられたデータのクラスタリングを行う．

本研究のアルゴリズム

入力: 情報表.

出力: 作成したクラスタの集合.

(i) 情報表から全ての縮約を求める．

(ii) 求めた全ての縮約から，各属性の出現率を計算する．

(iii) 出現率の閾値を設定する．

(iv) 出現率が閾値以下である属性を抽出する．

(v) (iv) で抽出した属性に基づく識別不能関係による同値類をクラスタとして出力する．

4 実験

前節で提案した手法を，サンプル数 101，属性数 17 である UCI Machine Learning Repository の Zoo データ [1] に適用した．17 個の属性のうちの 1 つである”type”を決定属性とし，得られた決定クラスを Zoo データにおけるクラスタとした．各クラスタは，哺乳類，爬虫類などの動物の分類に該当する．実験では，”type”を除く 16 個の属性からなる情報表に対して提案手法を適応した．

また，実験で得られたクラスタリング結果の妥当性を以下の式で評価する [2].

$$v_{X_i}(C_j) = \min\left(\frac{|X_i \cap C_j|}{|X_i|}, \frac{|X_i \cap C_j|}{|C_j|}\right).$$

ここで，X_i は本研究の手法により生成されたクラスタ，C_j は元のデータのクラスタであり，X_i の C_j に対する妥当性を表す.「X_i ならば C_j」という決定ルールを考えると，前者の式を確信度，後者の式を被覆度と見なすことができる．全ての X_i と C_j について妥当性の値を求め，その平均値をクラスタリングの妥当性の評価値とする．但し，X_i ならば C_j である関係が全く無い (妥当性の値が 0) 場合は，除いている．この妥当性の評価が高いほど，適したクラスタリングが行われていると見なすことができる．

閾値は 0，0.2 および縮約に含まれる属性の出現率の平均値 0.6875 の 3 種類を用いた．それぞれの閾値での属性数，構成されたクラスタの個数，およびクラスタリング結果の妥当性の評価を表 1 に示す.

表 1: Zoo データに対して適用した結果

閾値	0	0.2	平均値
属性数	2	4	7
クラスタ数	3	6	11
妥当性	0.3750	0.7500	0.5385

表 1 の結果から，閾値が小さすぎる場合および大きすぎる場合は，クラスタ妥当性が低くなった．理由として，閾値が小さすぎる場合はクラスタ妥当性の確信度に当たる部分の値が低くなること，閾値が大きすぎる場合は被覆度の値が低くなることが考えられる．一方，閾値 0.2 の場合では，各クラスタの特徴を強く反映した属性である feather(羽の有無),fin(ヒレの有無),milk(哺乳類か否か),backbone(脊椎動物か否か) が抽出された．このことから，提案手法において閾値を適切に設定することで，データの特徴をある程度反映したクラスタリングが行えることが示唆される．

5 まとめ

本研究では，情報表として与えられたカテゴリカルなデータに対して，情報表の縮約を用いることで，類似性あるいは非類似性を表す測度を導入することなく，かつクラスタの個数をあらかじめ決めずにデータをクラスタリングする手法を提案した．今後の課題として，より複雑な規則性を持つデータでの実験，他のクラスタリング手法との比較が挙げられる．

参考文献

[1] Forsyth, R.: *Zoo Database*, UCI Machine Learning Repository (1990).

[2] 平野 章二，津本 周作: ラフクラスタリングによる医療データの類似化の試み，第 19 回人工知能学会全国大会講演論文集, 1F3-05 (2005).

[3] 宮本 定明: クラスター分析入門—ファジィクラスタリングの理論と応用—, 森北出版 (1999).

[4] 森 典彦, 田中 英夫, 井上 勝雄 (共編): ラフ集合と感性〜データからの知識獲得と推論〜, 海文堂出版 (2004).

[5] Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers (1991).

連絡先

工藤 康生
〒 050-8585 北海道室蘭市水元町 27-1
室蘭工業大学工学部情報工学科
TEL 0143-46-5469 FAX 0143-46-5499
E-mail: kudo@csse.muroran-it.ac.jp