シンボリックデータのクラスタリングに関する研究*

1X−04

青木 洋次朗　矢口 博之　市野 学†
東京電機大学§

1　はじめに

多変量のデータ解析手法の一つに，クラスタリング（クラスタ分析手法）が挙げられる．クラスタリングとは，与えられた対象から性質の似たサンプルを集め，クラスタ（＝集合）を形成し，対象の分類を行うものである．一般に，対象間の距離（類似性）尺度を定義し，その尺度に基づいて分類を行う．しかし，これらの尺度の対象は量的データのみ，あるいは質的データのみに限定されている．解析対象の実体をなるべく反映した表現で記述すると，数値と記号が混在したシンボリックデータとなる．ゆえに，シンボリックデータを扱うことのできる，より一般的な距離尺度を定義することが重要である．また，クラスタリングの過程において，新たに形成されたクラスタが，どのような特徴を根拠にしているのかを知ることができないという問題点が存在する．

そこで本稿では，量質混在の記述を許すシンボリックデータに対して適用可能な，特徴に基づく階層的クラスタリング手法を提案する．そのために必要となる距離尺度として，記述の一般性を考慮したサンプルの隣接関係に着目する．この尺度を利用することにより，量質混在のデータに対して適用可能となるだけでなく，クラスタ形成の根拠となる特徴を知ることができる．

2　シンボリックデータ

シンボリックデータとは，量的な特徴と質的な特徴が混在して記述されるデータのことである．量的な特徴としては，

（Ⅰ）連続値をとる量的特徴（身長，血圧など）

（Ⅱ）離散値をとる量的特徴（年齢，靴のサイズなど）

などであり，質的な特徴としては，

（Ⅲ）順序の入った質的特徴（学歴，年号など）

（Ⅳ）名義的特徴（血液型，性別など）

（Ⅴ）木構造をとる特徴（車のエンジン形式など）

などである．

（Ⅰ），（Ⅱ），（Ⅲ）の特徴は，範囲をもつデータ（閉区間）も許容する．（Ⅳ）の特徴は有限集合も特徴値として認めることとする．

*A study on the cluster analysis of symbolic data
†Yojiro Aoki, Hiroyuki Yaguchi, Manabu Ichino
§Tokyo Denki University

3　カルテシアンジョイン演算[1)]

E_{ik} を k 番目の特徴 F_k におけるサンプルの値とする．このとき，特徴 F_k におけるサンプル対 E_{ik}，E_{jk} のカルテシアンジョイン $E_{ik} \boxplus E_{jk}$ を以下のように定義する．

（1）F_k が量的特徴または順序の入った質的特徴である場合，閉区間

$$E_{ik} \boxplus E_{jk} = [\min(E_{ikL}, E_{jkL}), \max(E_{ikU}, E_{jkU})] \quad (1)$$

とする．ここで，E_{ikL}，E_{ikU} はそれぞれ特徴 F_k 上の閉区間 E_{ik} の最小値，最大値である．$\min(E_{ikL}, E_{jkL})$ は，E_{ikL}，E_{jkL} の小さい方の値，$\max(E_{ikU}, E_{jkU})$ は，E_{ikU}，E_{jkU} の大きい方の値を取る演算である．

（2）F_k が名義的特徴である場合，和集合

$$E_{ik} \boxplus E_{jk} = E_{ik} \cup E_{jk} \quad (2)$$

とする．

（3）構造的特徴である場合，$N(E_{ik})$ を特徴値 E_{ik} に含まれる端点に共通な直近の親ノードとして，

$$E_{ik} \boxplus E_{jk} = \begin{cases} E_{ik} \cup E_{jk} & (N(E_{ik}) = N(E_{jk})) \\ N(E_{ik} \cup E_{jk}) \text{ に接続する端点の集合} & (N(E_{ik}) \neq N(E_{jk})) \end{cases} \quad (3)$$

となる．ただし，任意の特徴値に対して，

$$E_{ik} \boxplus E_{ik} = E_{ik} \quad (4)$$

とする．

4　Generality[2)]

与えられた特徴の集合を，$F_0 = \{f_1, f_2, \cdots, f_d\}$ によって表し，これらによって記述されたサンプルの集合を $\Omega = \{\omega_k, k=1, 2, \cdots, N\}$ とする．

サンプル集合Ωにおいて，2つのサンプル ω_i，ω_j によって形成されるカルテシアンジョイン領域を想定する．そこに含まれる他のサンプルの数を，generality（記述の一般性）と呼び，n_{ij} で表す．範囲を持つサンプルの一部が，形成されたカルテシアン領域に含まれる場合，そのサンプルは generality として数えないものとする．つまり，範囲を持つサンプルの全体が，形成されたカルテシアン領域に全て含まれる場合のみ，そのサンプルを generality として数える．

また，各サンプル対 ω_i，ω_j の特徴集合 f_k に関する generality n_{ij} を並べた行列を generality matrix と定義する．

5　距離関数の定義

特徴集合F_0に含まれる特徴f_pのもと，generality=n（$0 \leqq n \leqq N-2$）のときの隣接行列を以下とする．

$$G_n(\Omega \mid f_p) = [g_n(f_p)_{ij}], \qquad (5)$$

$$n=1, 2, \cdots, N-2, \quad p=1, 2, \cdots, d$$

ただし，ω_iとω_jの特徴f_pに関するgenerality が n に等しいときは，$g_n(f_p)_{ij}=1$とし，その他のときは$g_n(f_p)_{ij}=0$とする．また，$g_n(f_p)_{ii}=0$とする．

このとき，特徴f_pに関するサンプルω_iとω_jの間の距離を，次のように定義する．

$$d_p(\omega_i, \omega_j \mid f_p) = \frac{1}{N-2} \sum_{n=1}^{N-2} (n \times g_n(f_p)_{ij}) \qquad (6)$$

$$i=1, 2, \cdots, N, \quad n=1, 2, \cdots, N-2, \quad p=1, 2, \cdots, d$$

（7）式を基に，サンプルω_iとω_j間の特徴集合F_0に関する距離を次のように定義する．

$$d(\omega_i, \omega_j \mid F_0) = \frac{1}{d} \sum_{p=1}^{d} d_p(\omega_i, \omega_j \mid F_0) \qquad (7)$$

$$0 \leqq d(\omega_i, \omega_j \mid F_0) \leqq 1$$

6　提案手法のクラスタリングアルゴリズム

（7）式により，シンボリックオブジェクト間に新たな距離尺度を定義した．以下に，新たな距離尺度を用いたクラスタリング手法のアルゴリズムを示す．ここで，サンプル対の融和とは，特徴毎のカルテシアンジョイン演算を行うものとする．

① 全てのサンプル対，全ての特徴について，generality matrix を求める．
② （6）式から，特徴 f_p に関する全てのサンプル対の距離を求める．
③ （7）式から，全てのサンプル対の距離を求める．
④ 距離関数より求めた値が最小となるサンプル対を融和する．
⑤ 融和したサンプル対について，②で求めた（6）式の値が最小となった特徴を，クラスタ形成の根拠となる特徴とする．
⑥ ①〜⑤を実行不能になるまで繰り返す．

7　電子計算機による提案手法の評価実験

人工データによる評価実験を行った．実験には，冗長な特徴を含めた7特徴，30サンプルを用いた．使用したデータの詳細を表1に示す．

表1　人工データの詳細

グループ	特徴F_1	特徴F_2	特徴F_3	特徴F_4 ～ F_7
グループA	平均　5 分散　2	平均　10 分散　3	平均　3 分散　0.2	平均　0 分散　5 正規乱数
グループB	平均　−5 分散　1.5	平均　0 分散　2	平均　−2 分散　12	

本稿提案手法による実験の結果をデンドログラムとして図1に示す．デンドログラム上部に示されている数字は，サンプルの融和回数である．デンドログラム下部に示されている特徴は，該当する融和回数で形成されたクラスタの根拠となる特徴である．

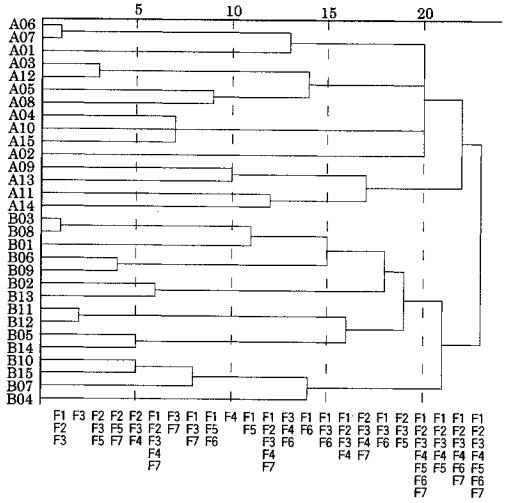

図1　実験結果（デンドログラム）

図1のデンドログラムから，グループAに属するサンプルのみから構成されるクラスタとグループBに属するサンプルのみから構成されるクラスタに分けることができた．ゆえに，未知のサンプル郡をグループA，グループBに分類することができたと言える．また，クラスタ形成の根拠となる特徴に注目すると，冗長な特徴に比して，特徴F1, F2, F3の出現頻度が高い．このことから，クラスタ形成の根拠となる特徴に基づいていると言える．

8　おわりに

記述の一般性を考慮したサンプルの隣接関係に着目する新たな距離尺度を定義した．その距離尺度を基に，シンボリックデータに適用可能な，特徴に基づくクラスタリングアルゴリズムを提案した．また，人工データによる評価実験を行ない，提案手法の有効性を確認した．

参考文献

1）矢口博之，市野学：「量質混在のデータに対する主成分分析の一般化」，進学論（A），Vol. J75-A, No. 10, pp. 1580-1589,（1992）
2）市野 学：「記述の一般性を考慮した近隣グラフの考察」，東京電機大学内部メモ（2001）
3）中川清太郎：「シンボリックデータに対するクラスタリング手法の研究」，東京電機大学理工学研究科修士論文（2001）