近隣グラフに基づく特徴分類*

3ZA－05

野水　俊明　矢口　博之　市野　学†

東京電機大学§

1　はじめに

データの中に存在するパターンを発見するための主要な方法として，クラスタリングや主成分分析といった方法が用いられている．しかし，サンプルを記述する特徴をどのように選択するかによって，解析結果に大きな影響を与える．したがって，複数の特徴間の関連を評価できる合理的方法が重要となる．

本研究では，新たな近隣グラフの概念である記述の一般性を考慮した近隣グラフ（Generality Ordered Relative Neighborhood Graph, GORNG）を導入し，近隣グラフを用いた類似性尺度を定義する．また一般的なデータであるシンボリックデータ（量的・質的な記述の混在したデータ）に対して適用できるクラスタリング手法を提案する．

2　カルテシアン・システムモデル[1]

特徴 F_k(k=1,2,…,d)によって記述される．5種類の特徴が本研究における解析の対象である．

1)連続値をとる量的特徴（重さ，長さ等）
2)離散値をとる量的特徴（年齢，家族等）
3)順序の入った質的特徴（年号，学歴等）
4)名義的特徴（性別，血液型等）
5)構造的特徴（エンジン形式等）

1)，2)，3)のとき閉区間，4)，5)のときは有限集合とする．

E_k を特徴 F_k の特徴値とすれば，サンプル E は特徴値の直積集合で(1)式で表される．

$$E=E_1\times E_2\times\cdots\times E_d \quad (1)$$

(1)式のように記述された直積集合を事象と呼ぶ．特徴 F_k の定義域を U_k で表す．定義域の直積空間を特徴空間と呼び，(2)式で表される．

$$U^{(d)}=U_1\times U_2\times\cdots\times U_d \quad (2)$$

2.1 カルテシアン・ジョイン演算

徴徴空間 $U^{(d)}$ 上のサンプル $A=A_1\times A_2\times\cdots\times A_d$ と，$B=B_1\times B_2\times\cdots\times B_d$ のカルテシアン・ジョイン $A\boxplus B$ は，各成分のカルテシアン・ジョインの直積は(3)式で定められる．

$$A\boxplus B=(A_1\boxplus B_1)\times(A_2\boxplus B_2)\times\cdots\times(A_d\boxplus B_d) \quad (3)$$

ここで $A_k\boxplus B_k$ は，以下のように定義される．

1)特徴 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，(4)式となる．

$$A_k\boxplus B_k=[\min(A_{kL},B_{kL}),\max(A_{kU},B_{kU})] \quad (4)$$

ここで A_{kL}，A_{kU} は，それぞれ閉区間 A_k の最小値，最大値であり，$\min(A_{kL},B_{kL})$，$\max(A_{kU},B_{kU})$ は，それぞれ A_{kL} と B_{kL} の小さいほうの値，A_{kU} と B_{kU} の大きいほうの値をとる演算である．

2)特徴 F_k が名義的特徴である場合，(5)式となる．

$$A_k\boxplus B_k=A_k\cup B_k \quad (5)$$

3)特徴 F_k が構造的特徴であるときは，$N(A_k)$ を特徴値 A_k に含まれる端点に共通な直近の親ノードとし，$N(A_k)=N(B_k)$ である場合，(6)式となる．

$$A_k\boxplus B_k=A_k\cup B_k \quad (6)$$

$N(Ak)\neq N(Bk)$ である場合，(7)式となる．

$A_k\boxplus B_k$={ノード$N(A_k\cup B_k)$に接続する全端点の集合} (7)

3　近隣グラフ GORNG[2]

データ自身の持つ幾何学的構造を表現する方法として，個々のサンプルの相対的な位置関係を基にした新たな近隣グラフを導入する．

ここで，特徴空間 $U^{(d)}$ 上で N 個の事象 E_k (k=1,2,…,N)が特徴 F_j (j=1,2,…,d)によって記述されていると仮定する．また，サンプルの集合を Ω とする．

3.1 Generality

任意のサンプル対 E_i と E_j の特徴集合 F に関するカルテシアン・ジョインが包含する他のサンプルの数を generality と呼び，(8)式で定義する．

$$\mathrm{gen}(E_i,E_j|F)=\left|\{E_k|E_k\in E_i\boxplus E_j\}\right| \quad (8)$$

$$0\le \mathrm{gen}(E_i,E_j|F)\le N-2 \quad k=1,2,\cdots,N \quad i\neq k,\ j\neq k$$

ただし，|・|は集合・の基数とする．

3.2 Generality matrix

各サンプル対 E_i と E_j の特徴集合 F に関する $\mathrm{gen}(E_i,E_j|F)$ を並べた行列を Generality matrix とし，(9)式であらわす

$$G(\Omega|F)=\begin{vmatrix} - & \mathrm{gen}(E_1,E_2|F) & \cdots & \mathrm{gen}(E_1,E_N|F)\\ \mathrm{gen}(E_2,E_1|F) & - & \cdots & \mathrm{gen}(E_2,E_N|F)\\ \cdots & \cdots & \cdots & \cdots\\ \mathrm{gen}(E_N,E_1|F) & \mathrm{gen}(E_N,E_2|F) & \cdots & - \end{vmatrix} \quad (9)$$

---

*Feature clustering based on the generality ordered relative neighborhood graph

†Toshiaki Nomizu , Hiroyuki Yaguchi , Manabu Ichino
§Tokyo Denki University

3.3 Generality Ordered Relative Neighborhood Graph (GORNG)

(9)式で定めた generality matrix から generality が n (0≦n≦N-2)の近隣グラフ(GORNG)を定めることができる.

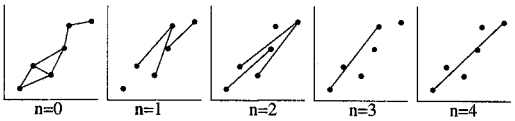

図1 サンプル数6の近隣グラフの例

特徴集合 F において, Generality が n の GORNG を(10)式のように定義する.

$$GORNG(n|F)=n \qquad (n \le N-2) \tag{10}$$

これらのグラフ群は Generality に関して大きさの順序を基にしている概念である.

4 特徴間の類似性尺度

特徴空間 $U^{(d)}$ 上における特徴間の類似性は, 着目したそれぞれの特徴上で見たときのサンプル間に見られる近隣関係の類似性によって表される.

2特徴 F_1, F_2 のサンプルの近隣関係が同一であるとすると GORNG に, 以下の関係が成り立つ.

$$GORNG(0|F_1)=GORNG(0|F_2)$$
$$GORNG(1|F_1)=GORNG(1|F_2)$$
$$\vdots$$
$$GORNG(N-2|F_1)=GORNG(N-2|F_2)$$

このことは, あるサンプルと他のサンプルとの間の Generality がそれぞれの特徴上で見たとき同一であることを示す. よって Generality の差を用いて類似性尺度を定義する.

特徴集合 F_p と F_q の類似性尺度を(11)式で定義する.

$$D(F_p,F_q)=1-\frac{2}{N(N-1)(N-2)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left|gen(E_i,E_j|F_p)-gen(E_i,E_j|F_q)\right| \tag{11}$$

5 特徴分類の方法

類似性尺度を用いて, 特徴間の因果関係を発見できるクラスタリングの方法を提案する.

STEP1 全特徴群対で類似性尺度(11)式を求める.

STEP2 類似性尺度が最大の特徴群対を融和する.
ただし最大値が2つ以上の場合, 最大値になる特徴群対の特徴を1つに融和し,
$F=\{F_p,F_q,F_r\cdots\}$ とする.

STEP3 融和できる特徴群対がなくなるまで,
STEP1,STEP2 を繰り返す.

類似性尺度の高い（特徴の振る舞いが近い）ものから順に融和される. そのため, 振る舞いの近い特徴同士のクラスタの発見を容易に行うことができる.

6 評価実験

特徴 F_1～F_3(-1～1の一様乱数)と, 表1の関数構造(線形, 2次関数, 3次関数)をもった特徴 F_4～F_{15}, 10種類の値をとる名義的特徴 F_{16}(ランダム), 表2のような関係を持つ特徴 F_{17}～F_{19} の19特徴を用いて, クラスタリングを行う.（全て300サンプル）

表1 量的特徴データ

特徴	F_4	F_8	F_{12}
構造	LNR(F_1)	F_1^2	F_1^3
特徴	F_5～F_7	F_9～F_{11}	F_{13}～F_{15}
構造	LNR(F_1)+rand	F_1^2+rand	F_1^3+rand

LNR0：線形関数, rand：-0.1～0.1の一様乱数

表2 名義的特徴データ

特徴	F_{17}	F_{18}	F_{19}
関係	F_{16}と同値	F_{16}と86%同値	F_{16}と67%同値

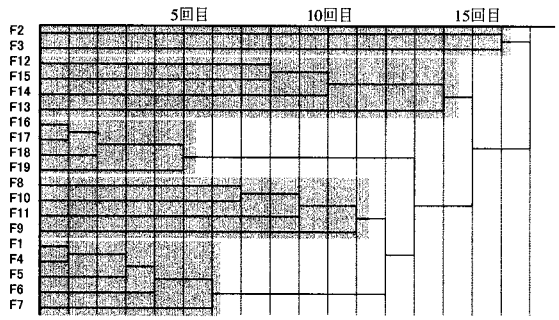

図2 結果のデンドログラム

図2のデンドログラムより特徴 F_1 に対しての構造別に融和されていることがわかる.

特徴 F_1 に対して線形の関係を持つ F_4 さらに F_5, F_6, F_7 と融和され, 特徴 F_{16} と同値の F_{17}, さらに F_{18}, F_{19} が融和され, 特徴 F_1 に対して2次関数の関係を持つ F_8, F_{10}, F_{11}, F_9 と融和され, 特徴 F_1 に対して3次関数の関係を持つ F_{12}, F_{15}, F_{14}, F_{13} と融和されていった. ただし, 特徴 F_5～F_7（F_9～F_{11} と F_{13}～F_{15} も同様）の融和順序は乱数を加えているため一定とは限らないが, 同一グループになっている. また, まったく関係のない乱数の特徴 F_2 と F_3 に着目すると, F_2, F_3 の意味のないものと残りの関係のある特徴群に大別することができる.

7 まとめ

サンプルの空間的分布が似ている特徴は, 類似性が高いという概念を基に新手法を提案した. また, 実験からシンボリックデータに対して, 似た性質を持つクラスタの発見に有用であるといえる.

参考文献

[1]市野学, 矢口博之："量的混在の記述を許す一般化されたミンコフスキーの距離", 電子情報通信学会 (A)J72-A,2,pp.398-405(1989)

[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)