7P-4

近隣集合に基づく特徴変数の類似性評価に関する研究

前野 良隆　　市野 学
東京電機大学 大学院 理工学研究科 情報社会学専攻

1．はじめに

シンボリック・データ・アナリシスは，データから有益な情報を獲得する統計手法の一般化を目指している．既存の多くの統計手法は，対象の記述が量的特徴に基づくことを前提としている．したがって，量的特徴でも区間を値とする場合や質的な特徴が混在する場合には，既存の統計手法の直接的な一般化は難しい．このような場合の手法として数量化理論が知られているが，量的特徴を区間に分けるという任意性の残る操作が必要であったり，数多くのダミー変数を用いることによる特徴の増大など，適用上対処すべき問題点を残している．

本研究では，カルテシアン・システム・モデルと呼ぶ数学モデルを用いる．この数学モデルでは，各対象が量的特徴と質的特徴の混在する形式で記述されることを認め，しかも特徴の型を統一することを目的とした特別な処理は必要としない．このカルテシアン・システム・モデルを基に，対象が記述される多次元空間において，対象間の相対的関係を記述する近隣集合の概念を定義する．次に近隣集合を用いて，多次元空間に対象群が構成する「幾何学的に薄い構造」を検出するための方法について考察する．また，提案手法の有用性を実験によって確認する．

2．相対近隣と近隣集合

2-1．相対近隣 (Nearest Relative Neighborhood)

量的・質的データを扱う事が出来る計算モデルとしてカルテシアン・システム・モデル（以下 CSM）[2]があり，相対近隣は CSM に基づく．CSM では，領域を定義するカルテシアン・ジョインという演算があり，サンプル対のジョインを作成した時，その領域内に含まれる他のサンプルの数 p を Generality p であるという．

d個の特徴F_x(x=1,2,…,d), n個のサンプルの有限集合$\Omega = \{\omega_1,\omega_2,\cdots,\omega_n\}$ によって記述されているとする．

特徴集合Fのp番目の特徴F_pにおいて, 任意のサンプル対ω_{ip},ω_{jp}によって形成される閉区間を想定する．その閉区間に包含されるサンプルをω_{kp}とし，ω_{kp}の集合ψ_{ijp}を，

$$\psi_{ijp}=\{\omega_{kp}|\min(\omega_{ip},\omega_{jp})\leqq\omega_{kp}\leqq\max(\omega_{ip},\omega_{jp})\} \quad (i\neq j,\ k\neq i,\ k\neq j) \qquad (2\text{-}1)$$

とする．ただし，$\min(\omega_{ip},\omega_{jp})$，$\max(\omega_{ip},\omega_{jp})$は，それぞれ$\omega_{ip}$と$\omega_{jp}$の小さい方の値，$\omega_{ip}$と$\omega_{jp}$の大きい方の値をとる演算である．

A study on the similarity of feature variables based on neighborhood sets
Yoshitaka MAENO , Manabu ICHINO
Department of Information and Arts , Tokyo Denki University

$$\text{generality}(\omega_i,\omega_j \mid F_p) = |\psi_{ijp}| \qquad (2\text{-}2)$$

ここで，ψ_{ijp}に含まれるサンプルの数を以下のように定義する．ただし，｜・｜は集合・の基数を表す．

注目する特徴F_pについて, 互いに隣接しているサンプル対，つまり，(2-2）式generality($\omega_i,\omega_j \mid F_p$)の値がkとなるサンプル対を，特徴$F_p$について k 相対近隣（k-Nearest Relative Neighborhood）であると定義する．

特に generality k が 0 の時を単に相対近隣と言う．

2-2．入れ子構造

入れ子構造とは，あるサンプル対の中に別のサンプル対が入った構造のことをいう．単調構造は最大サンプルと最小サンプルの中に内在する Generality が 1 個ずつ減少していく入れ子構造を有している．単調構造の量的特徴の例を図 2.1，質的特徴の例を表 2.2 に示す．

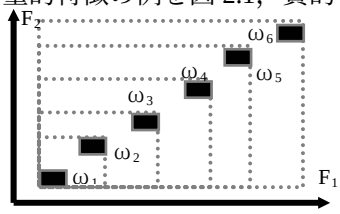

図 2.1　単調構造(量的特徴)

ω_1	a, b, c, d, e
ω_2	b, c, d, e, f
ω_3	c, d, e, f, g
ω_4	d, e, f, g, h
ω_5	e, f, g, h, I
ω_6	f, g, h, i, j

表 2.2　単調構造(質的特徴)

2-3．近隣集合 (Neighborhood set)

近隣集合は相対近隣に基づく．

近隣集合とは，ある基準となるサンプルの相対近隣の関係にあるサンプルω_{n-1}，ω_{n+1}と，基準となったサンプルω_nの集合であると定義する．

$$n(\omega \mid F) = \{\omega_{n-1},\omega_n,\omega_{n+1}\} \qquad (2\text{-}3)$$

例として図 2.1 を用いて説明する. 最初に 1 次元について考えたい. まずF_1についてサンプル集合の近隣を見ていくと，下記のようになる．

$$n(\omega_1 \mid F_1)=\{\omega_1,\omega_2\} \qquad n(\omega_2 \mid F_1)=\{\omega_1,\omega_2,\omega_3\}$$
$$n(\omega_3 \mid F_1)=\{\omega_2,\omega_3,\omega_4\} \qquad n(\omega_4 \mid F_1)=\{\omega_3,\omega_4,\omega_5\}$$
$$n(\omega_5 \mid F_1)=\{\omega_4,\omega_5,\omega_6\} \qquad n(\omega_6 \mid F_1)=\{\omega_5,\omega_6\}$$

F_1について見ると最初と最後が２つ, その間が自分を含む前後の３つずつになっているのがわかる． F_2，および，F_1 F_2の場合も同様の結果が得られる．このように，単調な構造とは全ての特徴において各サンプルの近隣集合が同じになるという特性を有する．

3．k 相対近隣を考慮に入れた単調性の程度の評価尺度

単調性の程度を評価する尺度として本提案手法は以下の式(3-1)で与える．この手法は近隣集合および近隣集合を generality k まで増やすことを許可したｋ相対近隣に基づく．

$$M = 1 - \left(\frac{1}{2\sum_{i=0}^{k}(2+k+i)+(3+2k)(n-2(k+1))} \sum_{k=1}^{n} \left| NRN_p(\omega_k \mid F) \right| \right) \times \frac{2k}{n-2}$$

$$0 \le M \le 1 \quad \cdots \quad (3\text{-}1)$$

ここで k はk相対近隣を示し，評価値は 1 をもっとも良い評価値とし，0〜1 の間に推移する．

4．性能評価実験

4-1．ノイズへの耐性

提案手法を評価するため，線形関数，指数関数，対数関数に逐次的にノイズを加えたデータを用いた．

各種関数を $v=f(u)$ によって記述されているとし，X, Y をそれぞれ独立な標準正規乱数とする．

線形構造では説明変数 u に標準正規乱数 X を発生させた時，ピアソンの積率相関係数を用いて目的変数 v を以下の式(4-1)で与える．

$$v = pX + \sqrt{1-p^2}Y \qquad (4\text{-}1)$$

v と u のピアソンの積率相関係数の値は p となる．

指数関数，対数関数では，説明変数 u に一様乱数 Z [-5,5]を発生させた時，目的変数 v を以下の式(4-2)で与える．

$$v = f(u) + \alpha Y \qquad (4\text{-}2)$$

ここでαは，ノイズの大きさを指定するパラメータであり，αは 0.0 から 0.1 刻みで 1.0 まで変化させたものである．

実験に使用したデータはいずれも，サンプル数は 100，特徴数は 12 である．特徴 1 を u とおき，特徴 2〜12 はαを 0.0 から 1.0 まで変化させたときの v である．

ノイズがない状態でのピアソンの積率相関係数との比較を図 4.1，ノイズを加えていったデータに対しての実験結果を図 4.2 に示す．

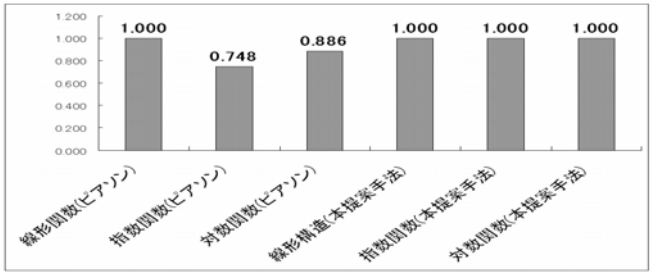

図 4.1　ピアソンの積率相関係数との比較

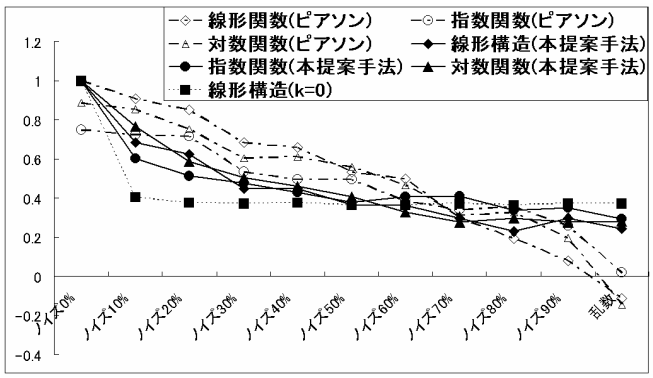

図 4.2　ノイズへの耐性

k相対近隣を考慮しない（k=0）時，ノイズのないデータには単調構造を抜き出すことが可能である．しかし，ノイズが少しでもある場合，単調構造を抜き出すことが不可能である．

次にk相対近隣を考慮に入れた本提案手法ではノイズが加わったデータに対してもノイズの量によって評価値が下がっており，ピアソンの積率相関係数に近い評価値を示すことができた．これはノイズの量に対して，本提案手法が適性に評価ができたと考える．

また指数関数，対数関数に関して，ピアソンの積率相関係数を用いた場合，ノイズのない場合でも約 0.750, 0.900 の値であったのに対し，本提案手法ではノイズのない場合，k相対近隣を考慮することによって 1.000 の値をとることが示された．またノイズに対しても，ノイズの量によって評価値が下がる傾向にあり，ノイズの量に依存した合理的な評価値が得られている．

4-2．質的な特徴を含む人工データ

提案手法を評価するため，質的な特徴を含む特徴集合によって形成される人工データを用いて評価実験を行った．実験に使用したデータは，サンプル数は 12，特徴数は 5 であり，質的特徴とする．特徴 4，5 を質的特徴の単調構造，残りの特徴は規則性を持たない冗長な特徴である．

本提案手法は CSM を用いているため，質的なデータであっても，単調な構造は抜き出すことが可能である．本提案手法の実験結果を図 4.3 に示す．

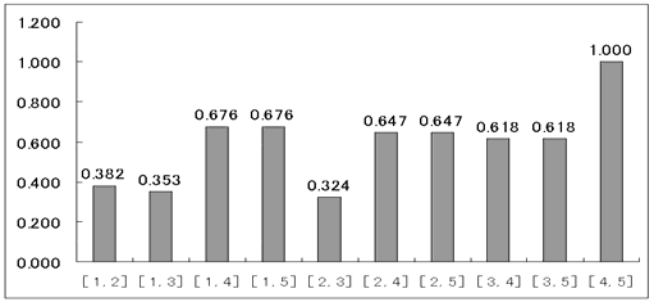

図 4.3　実験結果

実験の結果，本提案手法は質的な特徴が内在しているデータであっても評価することができた．また冗長な特徴が含まれたデータに対しても，単調な構造を抜き出すことが可能である．この結果，量的特徴，質的特徴の混在したシンボリックデータに対応していることが実証できた．

5．まとめ

本研究では特徴間の類似性の尺度を見るために近隣集合を提案した．またk相対近隣を用いることにより，k 相対近隣を考慮に入れた単調性の程度の評価尺度を定義した．簡単な人工データによる性能評価実験において，提案手法の有用性を示した．

参考文献

[1] M.Ichino , Detection of Monotonic Chain Structures in Mixed Feature Type Multidimensional Data ,2005.

[2] 市野学，矢口博之，野中武志，“幾何学的厚みに基づく相関係数”，電子情報通信学会論文誌A，Vol.J85-A，No.4，pp.490-494，(2002).