一般化された相関係数に関する研究*

3ZA－04

大阪 文洋† 矢口 博之 市野 学‡
東京電機大学§

要旨
プロフィールデータに内在する因果関係の検出は，データ解析の主要なテーマの一つである．特徴(属性，変数)間の因果関係を評価する尺度として相関係数が有用であるが，一般的な関数関係を扱うことはできない．そこで，本研究では非線形な関数関係を含む，一般的な因果関係の評価に適用可能な，一般化された相関係数について考察する．

1 はじめに

特徴間の関係を評価する代表的な指標に，ピアソンの積率相関係数(r で示す)がある[1]．積率相関係数は，平均値や分散，共分散といった基礎統計量を用いて特徴間の線形的関連の程度を評価可能である．しかし，図1に示すような特徴間の関係については，無相関(r=0)という結果を与える．図1(a)の関係は2次関数，(c)は円構造という明白な因果関係であり，(b)はなんらかの外部要因で制御される，3つの線形的関係の重ね合わせと見ることができる．このような特徴間の関係は線形的構造を含めて，「幾何学的に薄い」という共通の性質を有している．したがって，幾何学的な薄さを評価する合理的な仕組みを作れば，特徴間の因果関係を評価する尺度が構成可能になる．

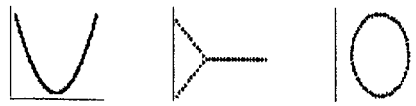

図1 幾何学的に薄い構造

幾何学的な薄さを評価する合理的な手法の一つとしてカルホーン相関係数(CAL で示す)がある[2]．カルホーン相関係数は，関数同定のような，サンプルの分布に適当な関数構造を当てはめるという過程を必要としない方法であり，特別な形をした領域に含まれるサンプルの数に関して定められる．このことが，幾何学的に薄い構造を評価可能としている．しかし，図1(c)に示す円構造のような，閉じた中空の構造は幾何学的に薄い構造であっても検出できていない．

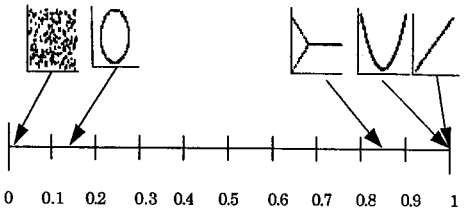

図2 カルホーン相関

本研究では，中空な構造も評価可能なより一般化された相関係数を定義する．

2 9つの領域

2次元ユークリッド平面上にN個のサンプル$\{x_1, x_2, \cdots, x_N\}$が与えられているとする．サンプル x_p と x_q で形成する9つの領域を図3のように定める(但し，$x_{p1} < x_{q1}$，$x_{p2} < x_{q2}$ とする)

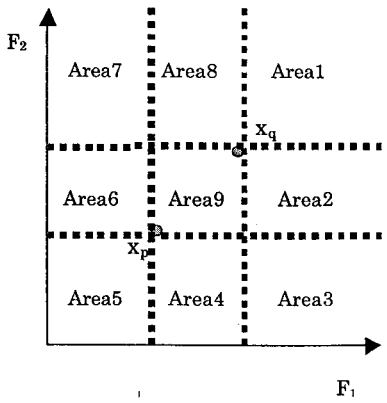

図3 サンプル対(x_p, x_q)で生成される9つの領域

サンプル対で生成される9つの領域に含まれるサンプル数をサンプル含有数とする．サンプル含有数を，

$$a_m(x_p, x_q) = |\{x_y \mid x_y \in area_m(x_p, x_q)\}| \quad m = 1,2,\cdots,9 \quad p \neq q \qquad (1)$$

但し$|\bullet|$は集合$\bullet$の基数とする
とする．

---

*A study on generalized correlation coefficient
† Fumihiro Osaka
‡ Hiroyuki Yaguchi, Manabu Ichino
§ Tokyo Denki University

3　分散の最小値を判断基準に用いた手法

分散最小値を判断基準に用いる手法(以下：VM_CAL) は,サンプル対(x_i,x_j)で生成される９つの領域に含まれるサンプル数の分散が最も小さくなるような x_i と x_j を選択する手法である.分散値が最小のときの各領域におけるサンプル含有数を式(3)を用いて算出する.

定式化すると以下のようになる.

$$s=\frac{n\sum_{k=1}^{n}a_k(x_i,x_j)^2-(\sum_{k=1}^{n}a_k(x_i,x_j)^2)^2}{n(n-1)} \qquad (2)$$

$$(i=1,2,\cdots,N-1\ ,\ j=i,i+1,i+2,\cdots,N\ ,\ k=1,2,\cdots,9m,i\neq j)$$

if　Min(s)　then

$$VM\_CAL=1-\frac{\text{各領域の下位4位までの平均}}{\text{各領域の上位4位までの平均}} \qquad (3)$$

これにより各領域に含有されるサンプル数の比は幾何学的に薄い構造ならば偏ったサンプル含有数となり,幾何学的に厚い構造ならば満遍なくサンプルが含有されると推測できる.

分析手順を以下に示す.

Step1　サンプル対(x_p, x_q)について,生成された９つの領域の分散 s を求める

Step2　step1 を N(N-1)/2回繰り返す(全てのサンプル対について行う)

Step3　分散 s が最小のとき、各領域に含有されたサンプル数を用いて式(8)を算出する

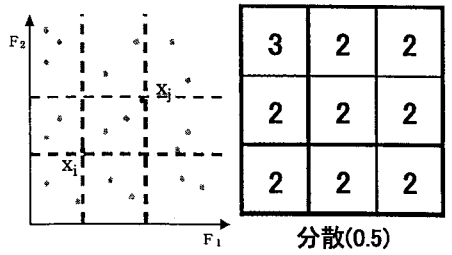

図 4.1　分析手順（評価値　0.12）

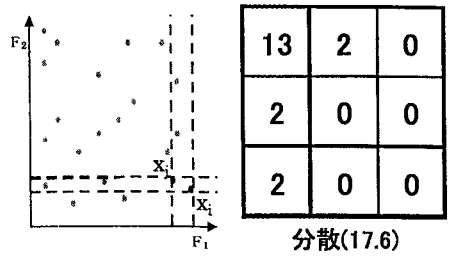

図 4.2　分析手順（評価値　1）

図 4.1 に示したように, 分散が最も小さいとき,各領域のサンプル含有数は理想的な比になる.

4　人工データによる実験

VM_CAL の性能を評価するために,図５に示す人工データによってカルホーン相関係数と比較実験を行なった.使用データは線形,２次曲線,乱数,円構造,Ｙ字型の１０特徴３２２サンプルを用いて行った.

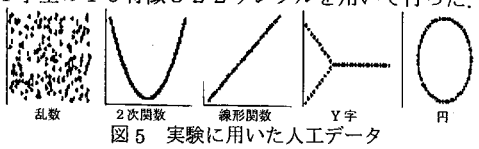

図５　実験に用いた人工データ

実験結果を図６に示す.

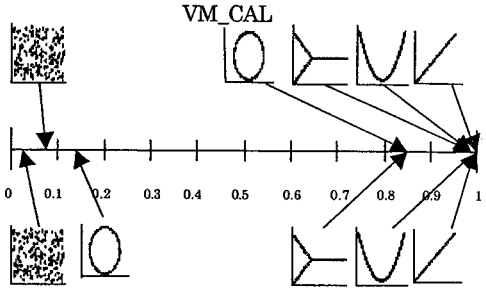

図６　実験結果

以上の結果から,VM_CAL はカルホーン相関では評価することのできなかった円のような中空な構造と線形構造や２次曲線などの幾何学的に薄い構造を同時に評価することができた.また,乱数のような幾何学的に厚い構造もカルホーン相関と比べると若干高い値となったが相関が無いと判断することのできる値におさまっている.

5　まとめ・今後の方針

VM_CAL を提案し,その有効性について検証した.その結果,幾何学的に薄い構造として,関数構造や低次の関数構造の重ね合わせ構造,円構造のような閉じた中空な構造等は、期待通りの結果を得る事ができた.

カルホーン理論ではサンプル数の３乗に比例して計算時間が増えるといった問題点が残されている.性能を落とさず処理能力を向上させることが今後の課題である.

参考文献

[1] S.S. Wilks, Mathematical Statistics, Wiley International Edition, 1962

[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”

[3] 市野 学 “カルテシアン・ジョイン・システムにおけるパターン認識” 信学会パターン認識理解研資, PRU-86-20

[4] 野中 武志　“相関係数の一般化に関する研究”　東京電機大学大学院理工学研究科システム工学専攻修士論文(2000)

[5] 野中, 木村　“カルテシアン・システム・モデルに基づく概念解析の研究”　東京電機大学理工学部経営工学科卒業論文(1999)