6C-3

近隣サンプルを利用した共変関係を有する特徴組の検出

名児耶　厚†　　石川　慎也†　　小野　裕次郎‡　　市野　学†
東京電機大学理工学部情報社会学科†　　十文字学園女子大学社会情報学部‡

1　はじめに

データ解析において，データに内在する特徴間の共変関係の検出は重要である．ピアソンの積率相関係数[1]は、特徴間の線形構造の評価を目的としている．さらに，より広い共変関係の評価法として、カルホーン相関係数[2]が提案されている．しかし、これらの方法は、基本的に 2 特徴間の関係に注目している．

我々は共変関係の検出対象を 3 特徴の場合にまで拡張し，より一般的な共変関係の評価手法確立を目指している．先の報告において，複数の特徴で同時に隣接関係を調べる MFRN: Multi Feature Relative Neighborhood というサンプル間の関係を用いて，2〜3 特徴の共変関係の評価に有効性が見られることを示した[3]．本報告では，共変関係を有する特徴組においては MFRN が鎖状に成立していくことに着目し，相対的に共変関係を評価する手法について提案する．

2　定義と用語の説明

2．1　幾何学的厚みと MFRN[3]

特徴間に共変関係が存在する場合，ある特徴において近隣にあるサンプル対は，他の特徴においても近隣に存在する傾向にある（図 1,2 参照）．さらに，この「幾何学的厚み」は 3 特徴の場合でも同様に保たれる．

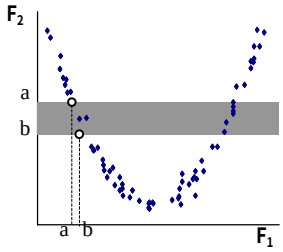

図 1　幾何学的に薄い構造

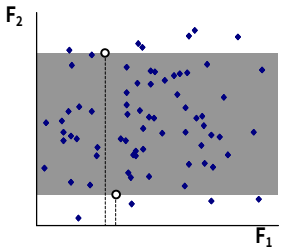

図 2　幾何学的に厚い構造

MFRN はサンプルの隣接性に着目することにより，幾何学的厚みの測定を行っている．与えられたデータが，N 個のサンプル ω_k(k=1,2,…,N)，d 個の特徴 F_k(k=1,2,…,d)によって記述されているとする． p 番目の特徴 F_p において，任意のサンプル対ω_i,ω_j によって形成される閉区間に含まれる他のサンプルの数を generality と呼ぶ．また，generality が 0 となるとき，これらのサンプル対ω_i, ω_j は，特徴 F_p に関して相対近隣(Relative Neighborhood, RN)であるという[4]．MFRN は 1 特徴における RN が複数の特徴で同時に成立している状態であり，線形構造など単調構造を有する特徴組に見られる特有の性質である．さらに，非線形構造など RN の同時成立が崩れてしまう場合でも，サンプル対の間に含まれるサンプル数のうち最大のものを許容度 p というパラメータとして保持し，その上での成立を認めることにより対処している．

2．2　データ構造と MFRN の関係

先の報告において，MFRN は

- データの構造に沿って成立していく
- 幾何学的に薄い構造の場合，狭い範囲の許容度に集中して成立していく

という性質を持つことを述べた[3]．本報告では前者の性質に着目し，共変関係の評価手法確立を目指す．

共通の終了条件を与えた場合，図 3 の例に示すように MFRN はデータ構造に沿った形で成立していく．

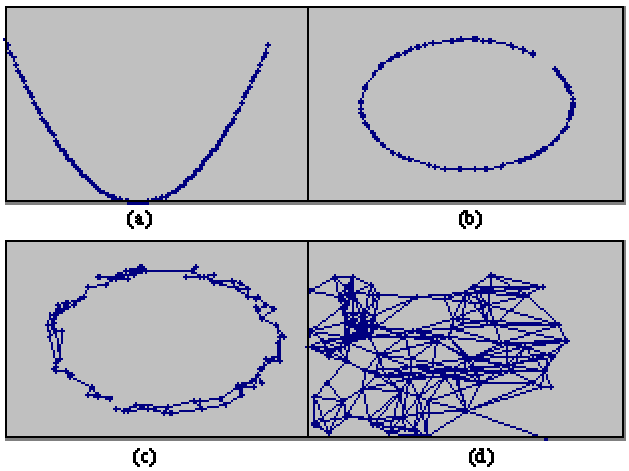

図 3　MFRN の成立の様子

図 3(a), (b)のように幾何学的に薄い構造のときは，構造の線形性，非線形性にかかわらず鎖状に成立

Detection of covariant relation in multi-dimensional data by using relative neighbors
†Atsushi Nagoya, Shinya Ishikawa, Manabu Ichino, Department of Information and Arts, Tokyo Denki University
‡Yujiro Ono, Department of Social Sciences and Information Science, Jumonji University

していき，(c)のようにノイズが混入するとその分乱れが生じる．(d)のように幾何学的に厚い場合は，データ空間全体に張り巡らされるように成立する．

２．３　評価方法

幾何学的に薄い構造のときには，MFRN が成立しているサンプル対（をグラフと見立てた場合）の位相距離と，そのサンプル対による generality とが（隣接状態を 0 とすると）同値であることに着目する．そこでまず，MFRN の成立している全てのサンプル対に対し，位相距離と generality との差を求める．次にその一覧から成る数値集合内で隣接する値同士の差（階差）を求め，それをサンプル対数で割って評価値とした．これは MFRN が 1 本の鎖状に成立しているときを下限とし，それからどれだけの乱れがあるかを数値として仮想的に評価している状態である．

この値はノイズがまったく混入していない人工データなどのときに，その構造にかかわらず 0 となり，ノイズの混入によって徐々に値が増加していく，非正規化評価値である．また，特徴組の特徴数が 2〜3 のときを区別せずに使えるため，多次元データに対して各々の特徴組の共変関係に関する順位付けをする，といった使用を想定している．

３　実験結果

３．１　様々な構造に対する実験

250 サンプルの人工データを用い，様々な構造，かつ 2〜3 特徴を混在させて評価値の違いを調べた．それぞれのデータにはノイズを加えていないため，乱数構造を除きほとんどが 0 に近い値となることを期待している．実験に用いたデータと実験結果を表 1 に示す．ただし，3 特徴のデータは 2 特徴のデータに 3 特徴目を追加して作成した．

表 1　データと実験結果

データ	特徴数	結果
乱数構造	2	24.25
線形構造	2	0
二次関数	2	0
三角関数	2	0
円構造	2	0
乱数構造	3	12.35
線形構造	3	0
二次関数	3	0
三角関数	3	0.04
円構造	3	0

表 1 より，幾何学的に薄い構造のデータでは全て 0 もしくはそれに近い値となり，乱数構造に対する明確な違いが見られる．

３．２　ノイズ付加に対する実験

表 1 に示したデータのうち，2 特徴の線形構造，三角関数に対して分散を徐々に増加させながら一様乱数を発生・加算して作成したデータを用い，ノイズの増加に対する評価値の変化を調べた．縦軸に評価値，横軸に分散値をとりグラフ化した結果を図 4 に示す．ただし，分散値はノイズなしのデータの値域を基準として，それに対する任意の割合という形式で生成した．

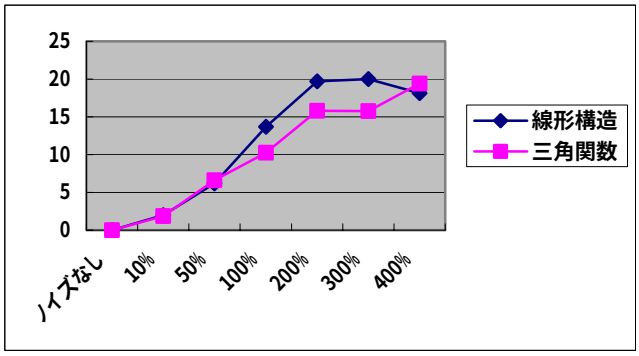

図 4　ノイズと評価値の変化

図 4 より，データの構造にかかわらず，ノイズの増加に対応して値が大きくなっていることがわかる．

４　おわりに

サンプルの隣接関係を利用した MFRN というサンプル対の関係を使い，2〜3 特徴の共変関係を有する特徴組を検出する方法を提案し，実験によりその性能を示した．今後の課題として，サンプル数・特徴数の影響による評価値の変動を取り除くことが挙げられる．

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”，電子情報通信学会論文誌，Vol.J85-A，No.4，pp.490-494，2002

3)名児耶厚，石川慎也，小野裕次郎，市野学，“サンプルの隣接関係に着目した多次元データに内在する共変関係検出に関する考察”，情報処理学会第 67 回全国大会，2004

4) Y. Ono, M. Ichino, “A New Feature Selection Method to Extract Functional Structures from Multidimensional Symbolic Data, IEICE Trans. Inf. & Syst., vol. E81-D, NO.6 June, 1998