2F2-04

多次元データにおける相関性の抽出

Finding the correlation between some features in multidimensional data.

松本　文士　　矢口　博之　　市野　学
Bunji Matsumoto　Hiroyuki Yaguchi　Manabu Ichino

東京電機大学　理工学部　情報社会学科
Tokyo Denki University, College of Science and Engneering, Department of Information and Arts

In this study, we define the Generality Ordered Relative Neighborhood Graph, which is the undirected graph being able to represent the information of sample position in the feature space and discuss feature selection for multidimensional data by the Geometrical Thickness.

1. はじめに

データから特徴間の相関性を見出す場合、ノンパラメトリック的手法ではモデルの選択という操作が必要となる。これは、モデルの適用の一意性が存在しないということを意味しており、この手法による相関性の抽出は不良設定問題そのものであると言える。

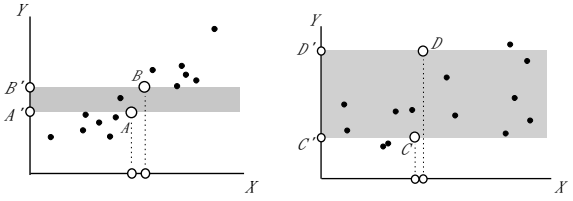
市野ら [1] は幾何学的厚みに着目することを提案している。これは特徴空間内でのサンプルの散らばりを空間的厚みとして捉えるもので、モデルの適用をおこなわず相関性の議論が可能であるという利点を持つ。

本研究では、特徴空間上のサンプルの空間的配置を Generality Ordered Relative Neighborhood Graph と呼ばれるグラフで表現し、多次元データにおける相関性を幾何学的厚みとして捉えられる可能性について議論する。

2. 幾何学的厚み

特徴間の相関性の有無は、一方では局所的、他方では大域的と、特徴空間上でのサンプルの拡がりに違いを与える。この差異を幾何学的厚みとして捉えることで、相関性の存在を議論することが可能である。

ある特徴軸上で近隣関係にあるサンプル対を他の特徴軸上に投射し、そのときにできる区間 (図 1(a) では $A^{'}B^{'}$, 図 1(b) では $C^{'}D^{'}$) を幅とした矩形領域を考える。すると、相関性がある場合（図 1(a)）の方が、そうでない場合（図 1(b)）よりも領域内のサンプル数が少ないことが分かる。このとき矩形領域の幅の違いから、前者を幾何学的に薄い構造、後者を幾何学的に厚い構造と呼ぶ。



(a) 幾何学的に薄い構造　(b) 幾何学的に厚い構造

図 1: 幾何学的厚み

3. Cartesian System Model(CSM)

CSM[2] は量質混在型のデータを直接扱うことが可能な数学モデルで *join* と *meet* の 2 つの演算子で構成されている。

連絡先: 松本　文士: E-mail matsu@csm.ia.dendai.ac.jp
東京電機大学　理工学部　情報社会学科
〒 350-0394　埼玉県比企郡鳩山町大字石坂

今、特徴集合 $\mathbf{F}$ によって形成される d 次元特徴空間 $\mathbf{U^{(d)}}$ を考える。$\mathbf{U^{(d)}}$ 上でのサンプル対 $(\mathbf{A},\mathbf{B})$ の *join* は次のように定義される。

$$\mathbf{A} \boxplus \mathbf{B} = (A_1 \boxplus B_1) \times (A_2 \boxplus B_2) \times \cdots (A_d \boxplus B_d) \quad (1)$$

このとき、$A_k \boxplus B_k$ は特徴 F_k での特徴値 A_k, B_k の *join* で、特徴 F_k が量的または順序の入った質的特徴のとき、

$$A_k \boxplus B_k = [min(A_{kL}, B_{kL}), max(A_{kM}, B_{kM})] \quad (2)$$

とする。A_{kL}, A_{kM} はそれぞれ閉区間 A_k の最小値、最大値である。また、特徴 F_k が質的特徴であるとき、

$$A_k \boxplus B_k = A_k \cup B_k \quad (3)$$

となる。もう一方の演算子である *meet* は次のように定義される。

$$\mathbf{A} \boxtimes \mathbf{B} = (A_1 \boxtimes B_1) \times (A_2 \boxtimes B_2) \times \cdots (A_d \boxtimes B_d) \quad (4)$$

このとき、特徴 F_k が量質問わず、

$$A_k \boxtimes B_k = A_k \cap B_k \quad (5)$$

である。

4. Generality

generality とは、*join* 演算によって空間上に作られる閉領域内のサンプル数を示す指標である。今、特徴集合 $\mathbf{F}$ でのサンプル集合 $\mathbf{\Omega}$ を以下の通りであるとする。

$$\mathbf{\Omega} = \{\omega_1, \omega_2, \cdots, \omega_N\} \quad (6)$$

このとき、サンプル対 (ω_i, ω_j) の *generality* を次のように定義する。

$$gen(\omega_i, \omega_j | \mathbf{F_a}) = |\{\omega_k | \omega_k \boxtimes (\omega_i \boxplus \omega_j) = \omega_k, for\ all\ k, k \neq i, j\}| \quad (7)$$

ただし、$\mathbf{F_a} \subseteq \mathbf{F}$ であるものとする。

5. Generality Ordered Relative Neighborhood Graph(GORNG)

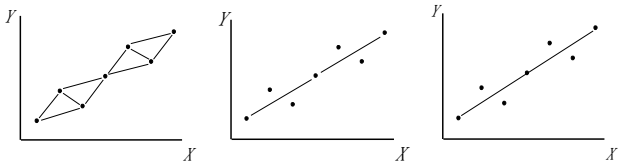
GORNG は特徴空間上でのサンプルの空間配置を表現することができる無向グラフで、以下のように定義される。

$$GORNG(n|\mathbf{F_a}) = (\mathbf{\Omega}, E(n|\mathbf{F_a})) \quad (8)$$

ただし、エッジ集合 $E(n|\mathbf{F_a})$ は以下の通りである。

$$E(n|\mathbf{F_a}) = \{e_{ij}^n | e_{ij}^n\ is\ the\ edge\ between\ \omega_i\ and\ \omega_j\ which\ are\ satisfying\ with\ gen(\omega_i, \omega_j | \mathbf{F_a}) = n\} \quad (9)$$

図 2 は *genearlity* が $0, 2, 5$ の GORNG の例を示したものである。



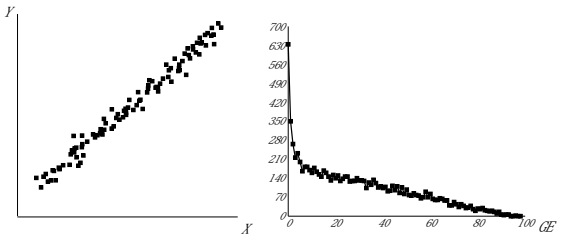
(a) $GORNG(0|X,Y)$ (b) $GORNG(2|X,Y)$ (c) $GORNG(5|X,Y)$

図 2: GORNG の例

6. GORNG 分布図

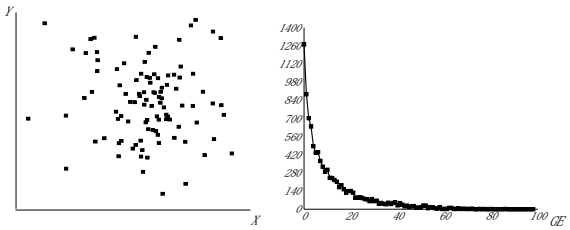
GORNG を用いることで容易に幾何学的厚みの差異を知ることが可能である [3]。図 3(b), 4(b) は薄い構造 (図 3(a))、厚い構造 (図 4(a)) を持つ人工データ（サンプル数は 100）に対して、*generality* が 0〜98 までの GORNG を施したときのエッジ数をまとめた GORNG 分布図と呼ばれるもので、横軸は generality、縦軸はエッジ数を示している。

この分布図から、薄い構造をもつデータに対しては、generality が増加するにつれてエッジ総数が全体として単調に減少しており、厚い構造では指数関数的な減少を示していることが分かる。よって、この分布図からデータの幾何学的な厚みを判別することが可能である。



(a) 薄い構造の人工データ (b) GORNG 分布図

図 3: 幾何学的に薄い構造と GORNG 分布図



(a) 厚い構造の人工データ (b) GORNG 分布図

図 4: 幾何学的厚い構造と GORNG 分布図

7. 多次元データにおける幾何学的厚み

ここで多次元データにおいても幾何学的厚みによる相関性の抽出が有効であるかの検討をおこなうために、サンプル数を一定にし、特徴数を増加させたときの薄い構造の存在の有無を調べた。

用いた人工データは、特徴対 (F_{2k-1}, F_{2k}) $(k = 1, 2, \cdots, d/2)$ のとき薄い構造、厚い構造の 2 つの構造が共に存在し、(F_{2k}, F_{2k+1}) $(k = 1, 2, \cdots, (d-1)/2)$ では、厚い構造のみを示すというデータである。ただし、2 つの構造が共存する特徴対での薄い構造に関与するサンプル数 P_{thin} と厚い構造でのサンプル数 P_{thick} は、

$$P_{thin} = N/\frac{d}{2} \quad (10)$$

$$P_{thick} = N - P_{thin} \quad (11)$$

であるとし、各サンプルはいずれか 1 つの特徴対でのみ薄い構造に関与するものとする。

また、薄い構造の存在は上述のように GORNG 分布図を用いて判断するのだが、データに観測ノイズが多分に含まれていると分布図から直接、単調減少区間を見出すことが難しくなる。そこで、あらかじめ適用されるモデルが明確になっているという理由から、分布図の *generality* 軸上での部分区間を設定し、その区間において回帰直線を適用したのち、その当てはまりの良さを示す指標である決定係数によって単調減少区間の存在の推定をおこなうこととした。

図 5 は計算機実験の結果で、横軸 FN は特徴数、縦軸 DC は決定係数を表わしており、サンプル総数が 200,300,400 のそれぞれの結果をグラフ化したものである。なお、決定係数は最も高い値を示した区間の値を用いている。この図から、特徴数が 8 まででは決定係数の値が約 0.8 以上と高い値を示しており、明確に薄い構造が存在していることが分かる。また、サンプル総数が増加することによって高い決定係数の値を示す特徴数の範囲が増加していることから、薄い構造の存在とサンプル数との依存関係を示しているものと考えられる。

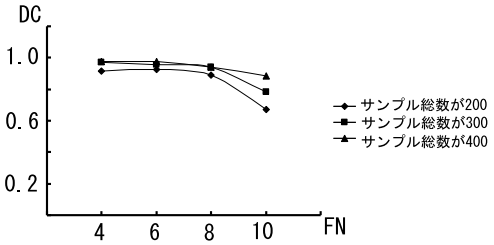


図 5: 特徴数と決定係数との関係

8. まとめ

多次元データにおける幾何学的厚さについて調べた。いくつかの異なった特徴対において異なった相関性が観られるという複雑さを持ったデータにもかかわらず、特徴数が 8 までなら安定して薄い構造を見出すことが可能であることを示した。また、このようなデータにおいても相関性の抽出が可能であるということは、複数個の局所レベルの相関性を、総体レベルの相関性として評価可能であることを示唆しているものと考えられる。

参考文献

[1] Y.Ono and M.Ichino: A new feature selection method to extract functional structures from multidimensional data, IEICE Trans. On Inform. And Systems, E81-D,6,pp.556-564,1998.Y

[2] M.Ichino and H.Yaguchi: Symbolic pattern classifiers based on the Cartesian system model, Data Science, Classification, and Related Methods, p.358-369, Springer, 1998

[3] 松本文士、市野学: 情報処理学会　第６４回（平成１４年）全国大会　3-227