グラフによる多次元データの構造解析

2W－7

森 康久仁　工藤 峰一　外山 淳　新保 勝
北海道大学大学院 工学研究科

1. はじめに

パターン認識で、多次元データの特徴的構造を解析することは非常に重要である。特に、各クラスが分離可能であるか、また、どの程度分布が重なっているかを知ることは識別子の設計において重要な情報となる。一般的に多次元データの構造を視覚化するには、何らかの基準で二次元あるいは三次元の低次元空間に射影する必要がある。しかし、クラスター構造や非線型構造などの構造を正確に表現することは難しい。そこで、本研究ではサンプルを囲う超区間によりクラスの構造をとらえ、グラフを用いてその構造を二次元表示し、解析する手法を提案する。

2. 従来の射影法

多次元データを二次元に射影する従来の方法は、大きく線形写像と非線型写像の二つに分類することができる [1]。本研究では代表的な線形写像である KL 展開と射影追跡法を用い、提案法との比較実験を行う。

2.1. KL 展開

確率ベクトル$\boldsymbol{x}$の共分散行列をΣとし、Σの固有値を$\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_r \geqq 0$、これらに対応する固有ベクトルを$\mu_1, \ldots, \mu_r$とする。この時、パターン$\boldsymbol{x}$の$\mu_1, \ldots, \mu_r$による展開

$$\boldsymbol{x} = y_1\mu_1 + \cdots + y_r\mu_r$$

を KL 展開と呼ぶ。ここでは、二次元に射影するためy_1, y_2を使う。

2.2. 射影追跡法

射影追跡 [2, 3] は多次元データのもつ興味深い構造を最も良く表すような、直線あるいは平面などの低次元空間への線形射影する方法である。興味ある構造を示す射影指標には様々のものが提案されているが、今回は、Friedman-Turkey の指標を用いた。

3. 部分クラス法

部分クラス法 [4, 5] では着目しているクラスに対するサンプルを正サンプルと呼び、それ以外のサンプルを負サンプルと呼ぶ。ここで、負サンプルを含まない極大の超区間を部分クラスと呼ぶ (図 1)。部分クラス法は各クラスを複数の部分クラスで表現する方法である。本研究ではこの部分クラスを用いて、データの構造を可視化することを考える。

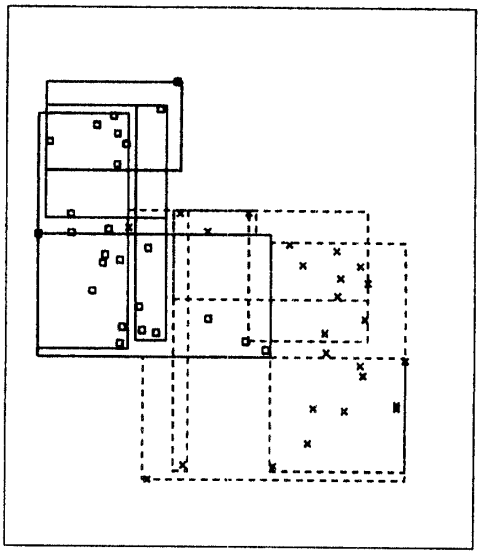

図 1: 二次元における部分クラスの例

(□と×はそれぞれのクラスのサンプルを、
実 (破) 線は部分クラスを表す。)

4. グラフによる表示

従来の表示方法では、与えられた特徴を何らかの方法で高次元から低次元に変換する。そのため、本質的に高次元距離情報を忠実に表現していない。そこで、本研究では前述の部分クラスを用いてデータそのものではなく、データ集合を表すと考えられる部分クラスをグラフ表現する。

部分クラスを$\{s_i\}(i = 1, \ldots, n)$、$w_{ij}$を$s_i$と$s_j$の共通部分の体積とする (全体が$[0,1]^n$に正規化されている)。ここで、グラフ$G = (V, E)$、$V = \{s_i, \ldots, s_n\}$、$E = \{(s_i, s_j) | w_{ij} > 0\}$を考える。次に、グラフ$G$の実際の二次元表示を与えるオペレータを$e$とする。これにより、一つの部分クラス$s_i$を

$$e : s_i \mapsto (x_i, y_i, w_{ii})$$

により、座標(x_i, y_i)に大きさw_{ii}で表現する。同様に、辺(s_i, s_j)に対しては

$$e : (s_i, s_j) \mapsto (e(s_i), e(s_j), w_{ij})$$

Structure Analysis of Multidimensional Data Using a Graph
Yasukuni Mori, Mineichi Kudo, Jun Toyama and Masaru Shimbo
yasu@huie.hokudai.ac.jp
Division of Systems and Information Engineering
Graduate School of Engineering
Hokkaido University, Sapporo 060, Japan

と表現し、$e(s_i)$ と $e(s_j)$ の間に太さ w_{ij}の辺を書く。ここで、(x_i,y_i) は全ての部分クラスの高次元中心位置の集合を KL 展開した。また、ある閾値以下の太さの辺は表示しないものとした。

5. 実験

10 次元の立方体の中に 10 次元の球を埋め込み、球の中と外でクラスが分かれているデータに対して実験を行った。主成分分析と射影追跡を施した結果を図 2 と図 3 に、部分クラス法を用いてグラフ表現した結果を図 4 に示す。図 2、図 3 において、+は球の内側、○は外側の点であり、図 4 において、○は球の内部の部分クラスであり、□は外側の部分クラスである。

実験結果から、従来の方法で表示した場合、10 次元のデータを無理に二次元で表示しようとしたため、データ構造をみてとることはできない反面、提案法においては、一方のクラスが近い範囲で多くの重なりを持って存在しており、他方のクラスは重なりの少ない、小さなクラスタを形成していることがわかる。

6. 結論

本稿では、新たな多次元データの構造表示方法を提案した。また、実際に多次元データをグラフ表現で視覚化した結果、従来の方法では適切にデータの構造をとらえる事ができなかった場合でも、データ構造を把握する事ができた。

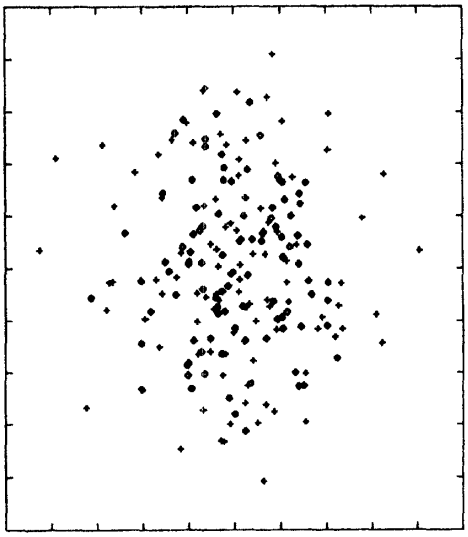

図 2: KL 展開による表示

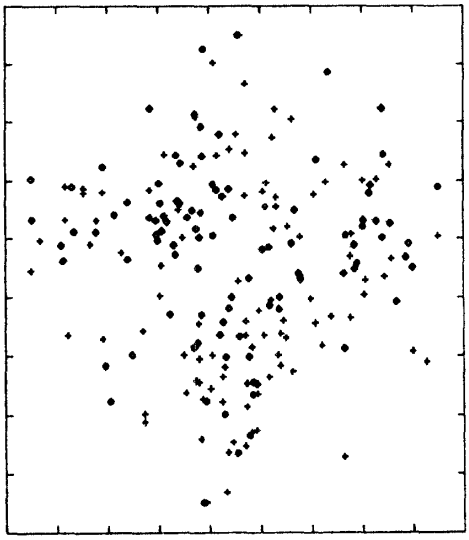

図 3: 射影追跡による表示

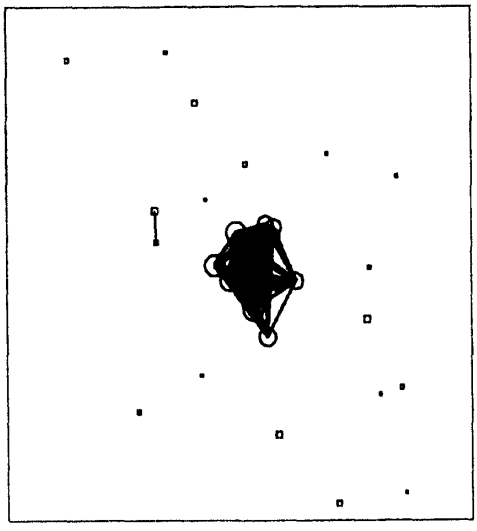

図 4: グラフによる表示

文献

[1] W. Siedlecki, K. Siedlecka and J. Sklansky, An Overview of Mapping Techniques for Exploratory Pattern Analysis. *Pattern Recognition*, **21**, 5(1988), 411–429.

[2] 小山一人, 射影追跡法の拡張とそれに基づくデータの非線型構造探索に関する研究. 北海道大学博士(工学) 学位請求論文, 1997.

[3] J. H. Friedman and J. W. Turkey, A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, **c-23**, 9(1974), 881–889.

[4] M. Kudo and M. Shimbo, Optimal Subclasses with Dichotomous Variables for Feature Selection and Discrimination. *IEEE Transactions on Systems, Man, and Cybernetics*, **19**, 5(1989), 1194–1199.

[5] M. Kudo and M. Shimbo, Analysis of the Structure of Classes and its Applications -Subclass Approach. *Current Topics in Pattern Recognition*, edited by Council of Scientific Information, India, 1989, 69–81.