# クラスタ判別法の医療データへの応用

3X－2

佐藤　新† 末永高志† 坂野　鋭†

†株式会社 NTT データ

## 1　はじめに

データからの知識発見や，パターン認識などで用いられる高次元のデータを低次元に写像し，人間に把握可能な散布図として表示すること，すなわちデータを可視化することは，データ解析の補助手段として極めて重要である．これまで我々は，強力なデータ可視化手法であるクラスタ判別法を提案し，既にいくつかのデータで有効性を実証してきた [1]．本稿においては，クラスタ判別法を医療データ解析に応用し，可視化手法としての有効性を確認するとともに，提案手法が異常値検出など，データ解析の上で有効な補助手段であることを実験的に示す．

## 2　クラスタ判別法

データマイニングで用いられるデータは通常数 10 から数 100 の属性を持つ高次元データであり，分布の構造を直感的に把握することは困難である．そのため，人間に理解可能な 2，3 次元の低次元空間に写像することが極めて重要であり，この目的のために主成分分析や自己組織化特徴写像などの方法が開発されてきた．しかしながらこれらの方法では高次元空間における分布構造の全てを保存しようとするため，結果的に構造を破壊してしまうようなゆがみを発生させることが避けられなかった．つまり，従来の可視化手法を用いた解析では，最初に可視化過程でデータの分布構造を破壊し，その上でデータの分布構造の解析を試みていいたことになる．

クラスタ判別法はこの問題を回避するために開発された手法である．その基本コンセプトは重要な構造情報を高次元空間で解析し，構造情報のみを保存することである．クラスタ判別法は，このコンセプトをクラスタ構造に着目することで実現した手法である．

従って，クラスタ判別法は，高次元空間でクラスタ構造を解析するステップとクラスタ構造を保存する写像によりデータを低次元に写像するための射影行列を計算するステップの 2 ステップで実現される．第 1 のクラスタ構造を解析する方法は従来よりクラスタ分析，もしくはクラスタリングとして知られる技術により実現される．第 2 のクラスタ構造を保存する写像は，写像された低次元空間でクラスタ内分散を最小化しクラスタ間分散を最大化する写像であると定義できる．この様な写像は一般に判別分析 [2] として知られるアルゴリズムによって実現される．以上の，クラスタリングと判別分析を組み合わせた可視化アルゴリズムを我々はクラスタ判別法と命名した．

**Cluster Discriminant Analysis and an application for medical data analysis**
Arata Sato†, Takashi Suenaga†and Hitoshi Sakano†
†NTT Data Corporation
Kayabacho Tower Bldg.,21-2, Shinkawa 1-chome, Chuo-ku, Tokyo 104-0033, Japan
{ara, suenaga, sakano}@rd.nttdata.co.jp

## 3　医療データへの適用

本節ではクラスタ判別法を医療データ解析に適用し，クラスタ判別法がデータ解析の補助手段として有用であることを実験的に示す．

一般に多次元データの解析ではデータの可視化が重要であることは前節までに強調しつづけたが，具体的には可視化によって分布構造の概略の把握や外れ値の検出が行われれば可視化はまず成功したと言えるであろう．また，検出された外れ値が統計的な外れ値であるのみならず当該分野における異常値であることが実証出来れば可視化技術は解析補助技術以上の意味を持つことになる．

以下，これらの観点で医療データの解析を試みる．実験に使用したデータは，髄膜炎の鑑別診断に関する医療データ [3] である．このデータは，患者 140 人分のレコード数を持つ．属性は 38 個あり，年齢，性別から白血球数，髄液細胞数等から構成される．

### 3.1　実験結果

本節では髄膜炎データのクラスタ判別法による解析過程を順を追って説明する．実験のために，クラスタリングアルゴリズムとして $k-$ 平均法 [4] を用い，判別分析としては Fishar の判別分析を採用した．

最初に図 1 にクラスタ数を 4 としたクラスタ判別法による可視化結果を示す．図中の丸で囲んだデータは，他のデータから離れており，少なくとも統計的な意味では外れ値であると考えられる．なお，以下に示す全ての図において，散布図に表れる同じ記号のデータは同じクラスタに属することを示す．

クラスタ判別法により，この様な外れ値が容易に検出できる理由は明らかである．外れ値は大多数を占める正常なデータから外れて存在するためにクラスタリングの段階で別のクラスタと判定され，かつ判別分析により正常データと離れた位置に写像される．無論，単

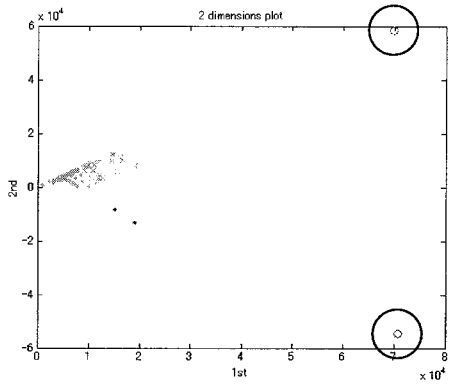


図 1: 可視化結果 1

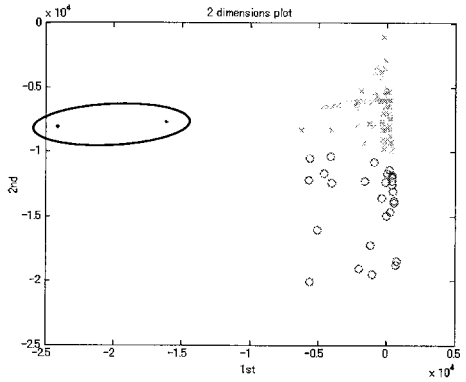


図 2: 異常値除去後の可視化結果 1

にクラスタリングにより得られた分割結果から外れ値を検出することは原理的には可能であるが全データに関する距離関係を解析する必要が生じてしまい，クラスタ判別法で与えられる可視化結果による直感的な検出に比して簡便性を欠くことは否めない．なお，この段階で検出された外れ値は白血球数が平均値より 2 桁程度多く，医学的にも異常であると推定される．

次に，これらのデータを取り除いた可視化結果を図 2 に示す．クラスタ数を 3 とした．この図でも丸で囲んだデータが他のデータよりも離れて分布している．これらのデータでは髄液に関係する属性で他のデータとは異なる傾向が見られた．

さらに，図 2 で検出された異常値と思われるデータを取り除き，クラスタ判別法を用いて可視化した結果を図 3 に示す．この図では，離れて分布するデータは存在しておらず，外れ値の検出が終了したと考えられる．

この最終結果では同じクラスタに属するデータ，すなわち高次元空間で近傍にあるデータが散布図の上でも近傍に存在することを示しており，良好な可視化結果を与えていると言える．この可視化結果を見る限り，湾曲構造などのあらわな非線形構造は検出できず，ルール抽出の上で多層パーセプトロンに代表される非線形識別技術を適用する理由が見出せないことがわかる．

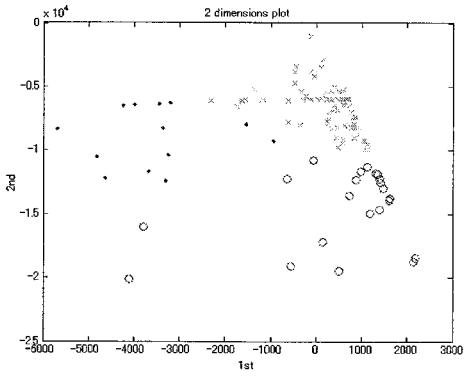


図 3: 異常値除去後の可視化結果 2

## 4 まとめと今後の課題

我々の提案する高次元データ可視化手法，クラスタ判別法を医療データに適用し，データの可視化及び統計的外れ値の検出などの観点でデータ解析への補助手段として有効であることを実験的に示した．

今後は今回検出した外れ値が医学的な観点で意味を持つ異常値であるか否かを検証するとともに，他の種類のデータに適用し，クラスタ判別法で摘出される外れ値が当該分野での異常値とどのような関係を持ち得るかを検証する．

## 参考文献

[1] 末永, 佐藤, 坂野, “分布の構造に着目した特徴空間の可視化-クラスタ判別法-”, 信学技法, PRMU2001-44, pp. 39-44, 2001.

[2] R. A. Fisher, “The use of multiple measurements in taxonomic problems”, Annals of Eugenics, Vol. 7, Part II, pp.179-188, 1936.

[3] 津本周作, “共通データに基づく知識発見手法の比較と評価”, 人口知能学会全国大会（第 12 回）論文集, pp. 72-73, 1998.

[4] J. MacQueen, “Some methods of classification and analysis of multivariate observations”, in Proceedings of the Fifth Berkeley Symposium on Mathemtical Statistics and Probability, pp. 281-297, 1967.