クラスタ判別法による外れ値検出

1X−05

佐藤　新† 末永高志† 坂野　鋭†
†株式会社 NTT データ

1 はじめに

データマイニングの分野において極端に大きな値や小さな値を発見すること，つまり外れ値検出は極めて重要である．外れ値がただのノイズであれば解析に悪影響を及ぼす．そのため，外れ値検出を行う様々な方法が提案されている．

しかし，外れ値に関しては統計学的な意味での客観的な基準を設けらないため，自動検出が困難である [1]．このため，現実には外れ値の検出は散布図の上で孤立しているデータを人間が主観的に外れ値とみなし行われる．しかしながらデータマイニングで解析の対象となるデータは通常数十から数百の属性を持つ高次元データであり，散布図を得ること自体が困難であった．

我々は高次元データの可視化手法としてクラスタ判別法 [2] を提案し有意な散布図が得られることを示してきた．クラスタ判別法は，高次元空間でクラスタ構造の解析を行い，クラスタ構造を保存する低次元空間への写像を求める可視化手法である．

本稿では，髄膜炎データの解析にクラスタ判別法を適用することで，外れ値検出に有効であることを実験的に示す．

2 クラスタ判別法による外れ値除去

クラスタ判別法は，最初にクラスタ高次元空間上でクラスタリングを行い，求められたクラスタを判別分析 [3] を用いて低次元空間への写像を求める可視化手法である．

我々はクラスタ判別法により外れ値が容易に検出できると考えた．というのは，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と考えられるためである．

無論，単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う．一方，クラスタ判別法を適用し可視化することにより，直感的な外れ値除去が簡便に行えるようになると考えられる．

Outlier detection using Cluster Discriminant 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 実験

本節ではクラスタ判別法を髄膜炎データの解析に適用し，解析に悪影響を及ぼす外れ値の検出に有用であることを実験的に示す．

高次元空間上で孤立しているデータが写像後の散布図上でも孤立していると人間が認識できれば，外れ値らしきデータが検出できる．さらにこのデータを除去し解析を行う．解析精度が良好であればこのデータは外れ値であり，外れ値の除去が出来たことになる．

今回の実験では，第1にクラスタ判別法を用いて高次元データの可視化を行い，外れ値らしきデータを人間が目で見て主観的に取り除く．第2にこのデータに対し，線形判別分析によりウイルス性と細菌性の識別問題に対する識別精度を求める．識別精度が向上すれば外れ値らしきデータは解析に悪影響を与える外れ値と判定でき，ノイズである外れ値の検出が成功したと言える．

実験に使用したデータは，KDD Challenge 2000 [5] で公開された髄膜炎の鑑別診断に関する医療データ [6] 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．

実験のために，クラスタリングアルゴリズムとして k-平均法を用い，判別分析のアルゴリズムとしてFisherの正準判別分析を採用した．

3.1 実験結果

クラスタ数を3としたクラスタ判別法による可視化結果を図1に示す．図中の丸で囲んだデータは他のデータから孤立しており，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．

外れ値除去を繰り返した場合のテストデータの識別精度と学習データ数を表1に示す．外れ値を除去し識

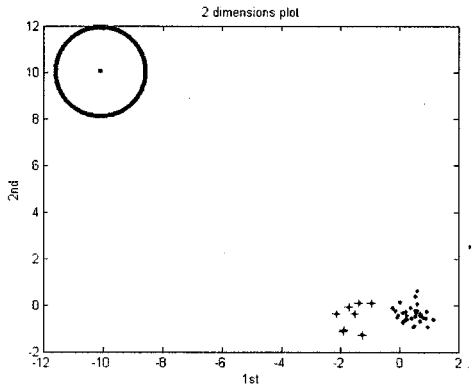

図 1: 外れ値除去回数 0

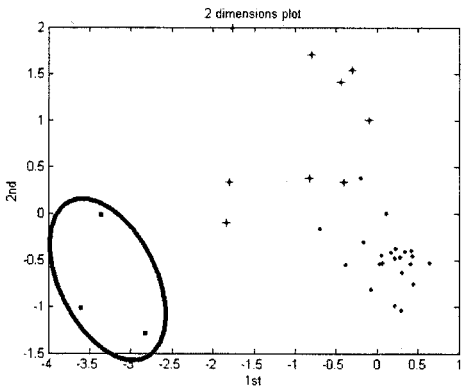

図 2: 外れ値除去回数 4

別精度が向上している場合，また学習データ数が少ないことを考慮に入れると識別精度に変化がない場合，除去したデータは解析に悪影響を及ぼしている外れ値であると考えられ，外れ値検出がうまく出来ていることが分かる．

5 回の外れ値除去を行った場合では，識別精度が低下している．外れ値除去を 4 回行った可視化結果を図 2 に示す．図中の丸で囲んだデータは外れ値か否か不明確であるが，このデータを除去した場合，識別精度は低下している．この結果より除去したデータが外れ値ではなく，ウイルス性と細菌性の識別に重要なデータであることが分かる．

このように，可視化結果から外れ値であると判断するためにはある程度の経験が必要となる．しかしながら，クラスタ判別法を用いることによりデータについて知識の無い人間による，外れ値検出が容易に出来るようになることが分かった．

表 1: テストデータの識別精度と学習データ数

外れ値	ウイルス性		細菌性	
除去回数	識別精度 (%)	データ数	識別精度 (%)	データ数
0	76.62	21	71.43	21
1	76.62	20	71.43	21
2	77.92	20	71.43	19
3	81.82	19	71.43	19
4	81.82	18	71.43	19
5	75.32	18	71.43	16

4 まとめと今後の課題

クラスタ判別法を髄膜炎データに適用し，外れ値検出の観点で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] 津本周作, “予兆発見と外れ値検出”, 人工知能学会研究会資料, SIG-FAI-A003-9, pp. 43-46, .2000

[2] 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.

[3] 石井健一郎, 上田修功, 前田英作, 村瀬洋, “わかりやすいパターン認識”, オーム社, 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.

[5] E. Suzuki (ed.), “Proc. Int'l Workshop of KDD Challenge on Real World Data”, http://www.slab.dnj.ynu.ac.jp/challenge2000.

[6] 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.