1. 異常検知
   1. 概要

異常検知とは多数のデータから外れ値等の異常なパターンを検出することである．例として，図 1に時系列データにおける異常なパターンを示す．

図 1　時系列データにおける異常の例[1]

異常検知には上述の例のようにある特定のデータが他のデータと比較して大きく外れた位置に存在する場合や，特定のデータではなく，データの振る舞いが他のものと異なる場合等を検出する手法である．

本案件では武繁・中澤モデルへの応用を目的とし，学習データとテストデータの性質に変化があった場合に異常検知を用いて検出する手法をまとめる．

一般的な異常検知の手法は以下の通りである．

* 距離を用いた手法
  + - 正規分布のデータを仮定する手法

個々のデータが独立に単一の正規分布に従うと仮定したうえで，新しく得られたデータの異常度を計算し，それが閾値以上であればそのデータが異常であると判断する手法．

* + - 非正規データを仮定する手法

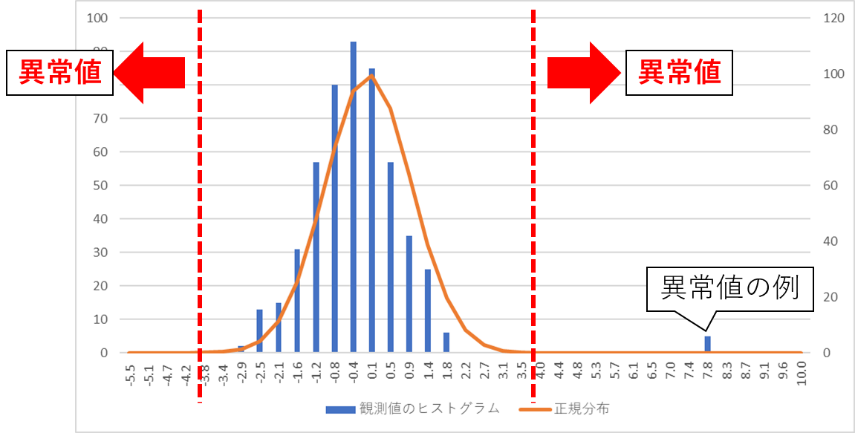
上述の正規分布を用いた方法において，ここのデータが独立に正規分布以外の分布に従うと仮定した場合の手法．

* 確率密度比を用いた手法
  + - 次元削減による手法

変数間に何らかの関係がある場合，次元削減により変数の座標とした多次元空間内の超平面を求め，その平面からの距離が閾値以上のデータが異常であると判断する手法．

* 1. 手法の詳細
     1. 正規分布を仮定した方法

本節では，正規分布を仮定した方法のうち，特に有名なホテリングの法と呼ばれる手法について記載する．当該手法では，観測値を正規分布とみなし，新たに得た観測値が正規分布の仮定の下で得られる可能性が低い場合に異常値とする．



一つのデータが次元ベクトルである個のデータを考える． ホテリングの法ではこのデータが，異常標本を含まないか，含んでいたとしても圧倒的に少数であるという前提のもとで各データが独立に以下の多変量正規分布に従うと仮定する．

|  |  |
| --- | --- |
| ：単一のデータ  ：平均  ：共分散行列 | (1) |

上記の分布に含まれるパラメータを決めるための自然な方法として，最尤法[[1]](#footnote-1)が挙げられる．この方法によると，パラメータの推計値は以下の式で与えられる．

|  |  |
| --- | --- |
|  | (2) |

次に，データの異常度を測る指標について述べる．異常度を測るデータを母集団分布からの独立標本を新たに観測したとすると，が得られる確率が低い場合は異常度が高く，確率が高い場合は異常度が低く算出されることが望ましいと考えられる．そのため，例えば以下の定義をすることができる[[2]](#footnote-2)．

|  |  |
| --- | --- |
|  | (3) |

(2)式を代入すると異常度の算式が導ける．ただし，観測データに関係ない定数は無視した．

|  |  |
| --- | --- |
|  | (4) |

異常検知のためには異常度の閾値が必要である．閾値を客観的な手法で決めることができれば，それを超える異常度をもつデータが異常であると判別することができる．

閾値を決定するため，異常度の確率分布を使用し，以下の算式に基づいて閾値を決める．

|  |  |
| --- | --- |
|  | (5) |

は誤報率と呼ばれ，比較的小さな値に設定することで閾値が意味のある量になる．例えばを3%とすると，が以上の値をとる確率が3%となり，以上の異常値を持つデータはデータ全体から見れば異常であるという解釈ができる．

なお，実用上，ほとんどの場合はが成立するが，この場合，は近似的に自由度，スケール因子1のカイ２乗分布に従うことが知られている[[3]](#footnote-3)．

|  |  |
| --- | --- |
|  | (6) |

以上より，当該手法をまとめると以下の通りである．

1. 予め決めているパーセント値に基づき，以下の方程式を満たすを算出

|  |  |
| --- | --- |
|  | (7) |

ここで，は自由度1，スケール咽ぢ1のカイ２乗分布である．

1. 正常標本が圧倒的多数を占めると信じられるデータから標本平均と標本分散共分散行列を求める．

|  |  |
| --- | --- |
|  | (8) |

1. 新たな観測値に対して以下の式で異常度を計算する．

|  |  |
| --- | --- |
|  | (9) |

1. であれば異常であるとする．
   * 1. 非正規データを仮定する方法

前節ではデータが正規分布に従うという仮定の下での異常検知の方法を記載したが，本節ではデータが正規分布ではとらえきれない特徴を持つ場合の異常検知の方法として近傍法やカーネル密度推定に基づく方法を記載する．

なお，本節では，次元データ個からなるを考え，新たに観測したの異常判定をする場合を考え，には異常標本が含まれていないか，含まれていたとしても圧倒的少数だと信じられるとする．

* + - 近傍法

を中心としたＭ次元の球を考える．この時，以下の二つの異常判定ができる．

1. 近傍法

球の半径を決めたとき，その級の中に入る標本の数がある基準値以下ならば，は異常であるとする．具体的には以下の手順で異常検知をする．

1. なんらかの方法で半径と近傍数の閾値を決める．
2. 新たな観測値に対して，半径の範囲に入る標本をから選ぶ．
3. その標本の数が閾値を下回ればは異常．
4. 近傍法

に近い順に個の標本を選んだ時，それらを囲む球の半径がある基準値以上ならは異常であるとする．具体的には以下の手順で異常検知をする．

1. 新たな観測値に対して，近傍となる標本をから選ぶ
2. それらを取り囲む最小の半径を求め，それがを上回ればは異常．

時系列データからの異常パターンの検出においては近傍法が主たる手法として使われる．また，特に高次元の場合にはデータセットが分布する全領域に妥当なやを決めるのは簡単ではなく，この点が実用上の困難となりうる．

* + - カーネル密度推定法

カーネル密度推定法は近傍法において，近傍に入るか否かという2値ではなく，着目する点からの距離に応じた判断に緩和することで使い勝手を向上させた手法である．

* + 1. 次元削減による方法
* 主成分分析による手法
* 確率的主成分分析による手法
* カーネル主成分分析による手法
  + 1. 時系列に対する異常検知の手法
       - 近傍法
       - 特異スペクトル変換法
       - 自己回帰モデルによる異常検知
       - 状態空間モデルによる異常検知

参考文献

1. 異常検知・変化検知

1. 観測したデータが得られる確率が最大になるようにパラメータを決める方法． [↑](#footnote-ref-1)
2. 情報理論の観点でもこの定義は筋の通ったものとなっている．ありふれたデータよりも，「珍しい」データ（＝異常値）を得る方が得られる情報量が大きいと考えれば，異常度が高いなら情報量が多く，異常度が低いなら情報量が低いと考えられる．情報理論においては，情報量は確率密度の対数値にを掛けた量で定義されるため，(3)式と整合的である． [↑](#footnote-ref-2)
3. が成立しない場合は，以下の統計量（ホテリングの統計量）が自由度の分布に従う． [↑](#footnote-ref-3)