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．

本報告書の構成は以下の通りである

* 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．

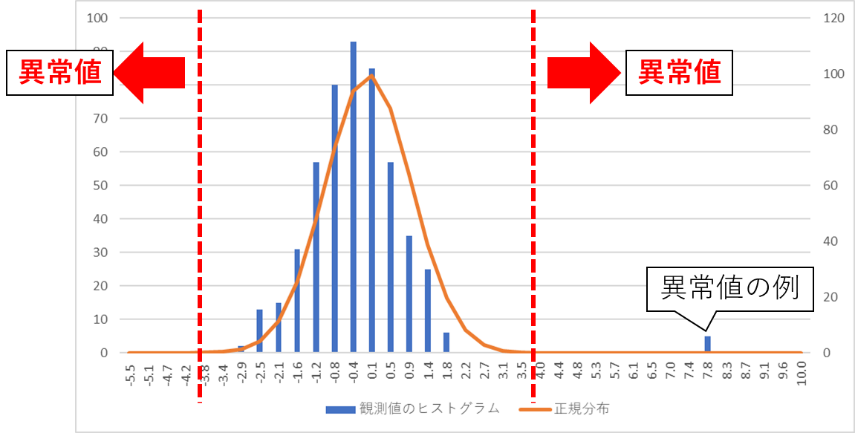


図 2　正規分布による異常検知のイメージ図

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) |

(4)式はマハラノビス距離と呼ばれる．この距離はデータの相関を考慮したものである．図 **3**はユークリッド距離とマハラノビス距離について，それぞれ距離が2である等高線を示したものである．分散が大きい軸については距離の尺度が小さく，分散が小さい軸については距離の尺度が大きいことがわかる．



図 3　ユークリッド距離とマハラノビス距離の違い

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) |

以上の手法を武繁・中澤モデルに応用する場合，以下の手順で進める事が考えられる．

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，スケール因子1のカイ２乗分布である．

1. 学習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し，パラメータを更新する．

複雑な系を考えた場合，次元数の値は数百から数千になることがある．この場合，上述の手法ではあるデータ自体が異常であると判定できても，そのデータを構成する各変数のうちどの変数が原因で異常となっているのかはわからない．以下で説明するマハラノビス=タグチ法では選択した変数に対応する異常度を計算することでこの課題に対応することができる[[4]](#footnote-4)．

当該手法では正常データが圧倒的多数だと信じられるデータセットと，異常と判明しているデータセットに対して以下の指標を計算する．

|  |  |
| --- | --- |
| ：選択した変数の集合  ：選択した変数の数  ：の分散共分散行列を使ったときの異常度 | (10) |

この指標はSN比と呼ばれる．上述の通り，変数の数が大きい場合，は期待値がのカイ2乗分布に従う．したがって，(10)の分子の1はの期待値を表すと考えることができる．SN比はこの1という値が正常値を表す基準値にとり，異常状態の1変数ああたりに異常度との比を計算していると考えられる．

例として，の場合を考える．この時，第変数が内のデータの異常度に大きく寄与しているとすると，真数の値が小さくなるため，SN比の値は大きくなる．したがって，SN比が大きくなるほどその変数の寄与度が大きいと考えられる．

* + 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．

近傍法ではk近傍法と近傍法の2種類が存在する．k近傍法はデータ点を中心とした球内のデータ数が基準値以下であれば異常なデータとみなし，近傍法は他のデータ点までの距離が基準値よりも大きい場合に異常なデータとみなす手法である．

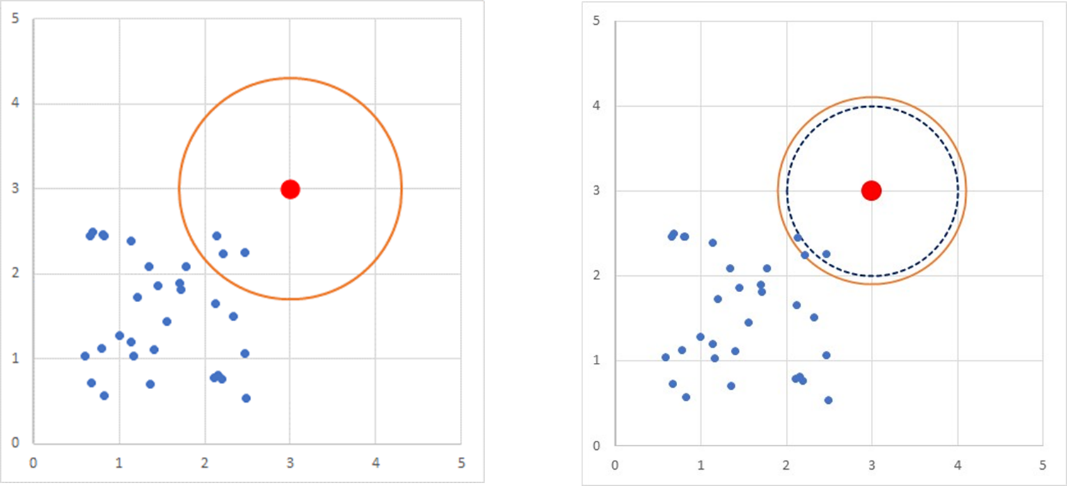


図 4　k近傍法と近傍法のイメージ図

以下では，次元データ個からなる武繁・中澤モデルの学習データを，新たに観測したデータの異常判定をする場合を考える．なお，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

を中心とした次元の球を考える．この時，二つの近傍法による異常判定の手順は以下の通りである．

1. 近傍法
2. なんらかの方法で半径と近傍数の閾値を決める．
3. 新たな観測値に対して，半径の範囲に入るデータの数が閾値を下回ればは異常あるとみなし，パラメータを更新する．
4. 近傍法
5. なんらかの手法で近傍数と半径の閾値を決める．
6. 新たな観測値に対して，個のデータを球内に含むような最小の半径が閾値を上回れば，は異常あるとみなし，パラメータを更新する．

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．

使用する距離としては，ユークリッド距離のほかにマハラノビス距離が使われる．

* + - 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．

まず初めに，観測値とに含まれる観測値との間の類似度Ｋを導入し，の確率密度を以下のようにモデル化する．

|  |  |
| --- | --- |
|  | (11) |

ここで，は類似度の到達距離を表すパラメータである．はを中心としてその周りの点にどれだけ影響を与えるかを表しており，一般に核関数またはカーネル関数と呼ばれる．

カーネル密度推定による異常検知では，まず初めにの推計を行う．基本的な手法としては以下の積分二乗誤差を最小化するように決める方法である．

|  |  |
| --- | --- |
| :真の確率密度 | (12) |

真の確率密度は実際には未知なので何らかの形で近似する必要がある．通常は以下のように近似することが多い．

|  |  |
| --- | --- |
| :真の確率密度 | (13) |

ここで，に含まれる式の内，未知のパラメータに依存しない部分は(定数)と記載した．

カーネル密度関数としてよく用いられるのは相関がない多変量正規分布である．

|  |  |
| --- | --- |
|  | (14) |

この場合，の値は以下のように計算される．

|  |  |
| --- | --- |
|  | (15) |

以上の手法を武繁・中澤モデルに応用する場合，以下の手順で進める事が考えられる．

* 1. (15)を最小化するバンド幅を推計する．
  2. 推計したバンド幅を(11)に代入し，以下の算式よりデータの異常度を計測する．

|  |  |
| --- | --- |
|  | (15) |

* 1. が閾値以上であれば，異常値とみなす．

参考文献

1. 異常検知・変化検知
2. Taguchi，G．and Jugulum，R.: The Mahalanobis-Taguchi Strategy: A Pattern Technology System，Wiley（2002）

メモ（報告書には載せない）

という値が式中に現れることからわかる通り，が平均から離れると異常度は大きくなり直感的な異常度と整合的である．また，については各座標軸を標準偏差で割ることを意味している．これは，データをある一つの座標上の位置で見たとき，例え平均から離れていてもその座標上の標準偏差が大きければ異常とはみなさないことを意味する．これを具体例で示したものがである．各データの縦軸の値は横軸の値より分散が小さい．例えば赤色のデータについて平均からの乖離という視点のみで異常度を判断すると，横軸の値は平均からの乖離が大きいため，異常と判断される可能性がある．しかしながら実際は異常といえるほどの乖離はない．これは横軸の値が平均から乖離していても，データのばらつき（＝標準偏差）が大きいためである．したがって，各軸の値を標準偏差で割り，ばらつきを揃える必要がある．



図 5　2次元データのイメージ図

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)
4. マハラノビス=タグチ法には複数の方法があるが，ここでは[2]の手法を記載する． [↑](#footnote-ref-4)