Technical Sheet
No. 21-26

データ解析入門 4 <T^2 統計量と Q 統計量を用いた異常検知>

キーワード：異常検知、主成分分析、T^2 統計量、Q 統計量、多変量解析

はじめに

品質管理や化学プラントの安全運転などにおいて、工場の各種センサー情報などを解析し、異常検知を実行することは重要です。本稿では T^2 統計量および Q 統計量を用いた異常検知手法について紹介します。

単変量データの外れ値

外れ値とは、ほかの数値から極端に逸脱した値のことです。外れ値のうち、逸脱の原因が明らかであるものを異常値といいますが、本稿では両者を区別せずに議論を進めます。

まず、観測対象が単変量データである場合を考えます。この場合、外れ値の確認にはヒストグラムや箱ひげ図などによる図示が有効です。箱ひげ図はヒストグラムを簡略化したグラフであり、数値の大小の序列から作成します（図 1）。また、箱ひげ図からは第 1 四分位点や中央値などを読み取ることができます。図中の 3 つの赤丸は所定の条件を満たすため外れ値に該当します。ただし、外れ値とみなされたとしても、データの棄却については慎重に判断する必要があります。

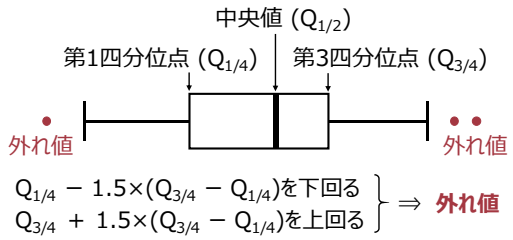

図 1　箱ひげ図

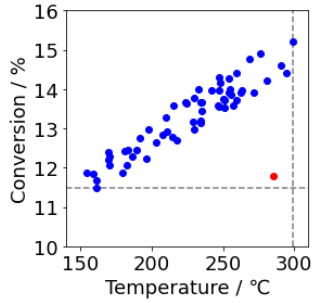

図 2　反応温度と転化率に関する仮想データ

多変量データの外れ値

単変量データとは異なり、データが多変量である場合は各変数をひとつずつ確認するだけでは外れ値を見逃してしまうことがあります。

簡単な例として、反応温度を高めると転化率が増大する仮想的な化学反応を考えます（図 2）。図 2 に示す通り、反応温度と転化率は直線的な関係にあり、その直線関係から逸脱したデータは異常であると判断できます。

しかしながら、異常データの反応温度と転化率はいずれも正常データの変域に収まっており、個々の変数の箱ひげ図を確認しても外れ値に該当するデータはありません。つまり、多変量データでは、変数間の関係性を無視すると外れ値を見逃してしまう恐れがあります。また、変数が多くなるとデータを図示することができません。そのため、化学プラントなどの多変量を扱う現場では、多変量解析や機械学習を用いた異常検知手法が重要になります。

主成分分析について

上述したように、多変量データの異常検知では変数間の関係を考慮することが重要です。変数間の線形関係をとらえ、少ない変数に情報を縮約する手法として主成分分析（PCA）があります。これまでに PCA による多変量データの可視化などについて紹介しています [1]。データの可視化に PCA を用いる場合は 2 つあるいは 3 つの主成分を使用することが一般的です。一方、異常検知のための主成分の適切な数は問題によって異なり、4 つ以上の場合もあります。後述の手法では、異常検出力や累積寄与率などを確認し、主成分の数を調整します。

異常検知の流れ

初めに今回の異常検知手法の手順を示します。

0)　正常データの収集
1)　PCA の実行
2)　異常検知に用いる主成分の数の設定
3)　T^2 統計量と Q 統計量の算出
4)　T^2 統計量と Q 統計量のしきい値の設定

地方独立行政法人
大阪産業技術研究所　本部・和泉センター
https://orist.jp/
〒594-1157　和泉市あゆみ野２丁目７番１号
Phone: 0725-51-2525（総合受付）

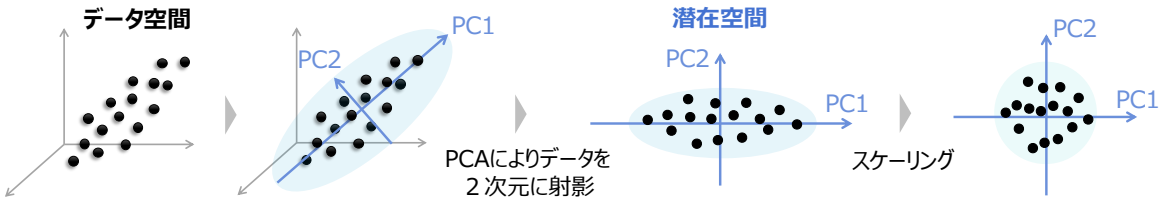

図 3　PCA を用いたデータ処理の流れ

実問題において、異常データは正常データに比べて圧倒的に少ないことがほとんどです。そのため、実用上、正常データのみを用いた異常検知手法が求められます。そこで、正常データと異常データとでは、計算される統計量は異なると仮定し、正常データのみを用いて異常検知を行います。今回は T^2 統計量と Q 統計量を用います。

図 3 に第 1 および第 2 主成分（PC1、PC2）を用いて T^2 統計量を求める流れを示します。まず、多変量データを PCA により次元削減し、主成分スコアを算出します。次に、各主成分スコアをそれぞれの標準偏差で割り、スケールをそろえます。このとき、各データ点と原点との距離の 2 乗が T^2 統計量です。

図 3 では高次元データを PC1 と PC2 が張る平面に写像しています。ここで、元のデータ点と 2 次元平面から逆写像した点との距離の 2 乗を Q 統計量とし、残差の影響を反映した統計量と考えます。両統計量に正常／異常判定のしきい値を設け、多変量データの主要な変動を T^2 統計量、残差による変動を Q 統計量で管理します（図 4）。

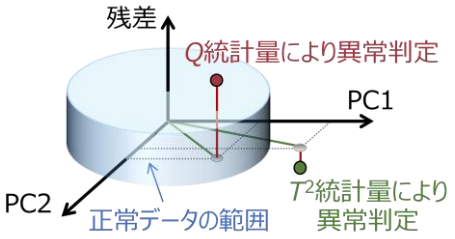

図 4　異常検知の概念図

異常検知の実行

本稿では、形状の類似した薬剤の近赤外（NIR）分光スペクトル [2]を用いて異常検知モデルを構築します。まず、NIR スペクトルに PCA を実行し、寄与率と累積寄与率を確認します（図 5）。次に、主要変動と残差を切り分ける主成分数を決定します。異常検知対象が特定されている場合、異常検出力を確認しながら主成分数を決定しますが、不特定の異常を広く検出したい場合は主成分スコアの分散や累積寄与率などを参考にします。ここでは第 2 主成分までを主要変動とし、両統計量のしきい値（破線）を 99 %点とした例を示

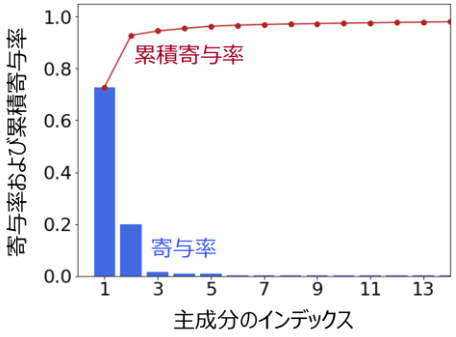

図 5　寄与率および累積寄与率

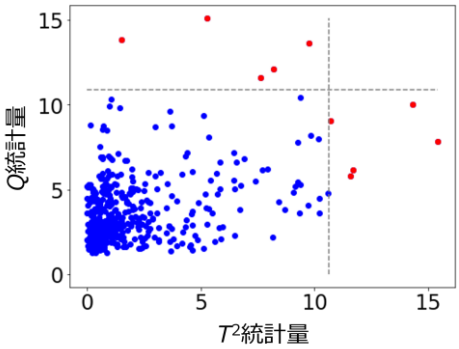

図 6　異常検知の例（青：正常、赤：異常）

しています（図 6）。本手法の適用例として、定常的な生産工程管理などが考えられますが、PCA によりデータの特徴をとらえられることを前提にしている点には注意が必要です。

おわりに

本稿で取り上げた異常検知手法は、多変量統計的プロセス管理手法として知られています。なお、Q 統計量がしきい値を上回り、異常と判定された場合、各変数の Q 統計量への寄与をプロットすることで異常の原因診断を行うことができます。

参考文献

1) 永廣卓哉 「データ解析入門 2 <主成分分析によるデータの可視化>」, ORIST テクニカルシート, No. 21-24 (2021)

2) Eigenvector Research, Inc. Homepage, https://eigenvector.com/wp-content/uploads/2019/06/nir_shootout_2002.mat_.zip（accessed on 26th May, 2021）

発行日　2021 年 12 月 20 日
作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@tri-osaka.jp