特集　正しく評価する～各分野におけるデータ解析処理～

昨今のコンピュータや分析機器などの発展により，得られるデータ量が一昔前と比べ圧倒的に多くなっている。さらに，インターネット上のデータベースから統計資料に至るまで，世の中には膨大な量の情報やデータが存在している。“データが出た”だけで満足してしまう技術者，研究者は（少なくとも本誌読者の中に）いないであろう。そのため，それらを“正しく評価する”ことがより一層強く求められている。例えば，工業分野ではハイスループット化による大量の製品の品質評価や化学プラントにおけるプロセス制御，生物科学分野ではDNAマイクロアレイ解析による遺伝子診断など，様々な分野においてデータ解析処理が必要である。さらには，地球規模での物質循環および測定データを基にした気象や災害の予測，治療や診断記録などの医療情報を活用した適切な看護や治療などにも必要不可欠であり，その活用範囲は多岐にわたっている。

そこで，本特集では各分野（化学工学分野を含む）におけるデータ解析処理手法について紹介し，さらにそこから得られる情報の意義や応用例などについても焦点をあてる。

（編集担当：宮永一彦）†

数値データの解析処理とプロセスシステム工学

山下 善之

1. はじめに

データ解析とは，実験や観測などによって取得したデータから，情報や知識を得るためにおこなう解析である。大規模複雑なデータの解析は，手法の研究分野や応用分野によって，統計的解析やデータマイニングなどと呼ばれているが，内容的には重なる部分も多く，明確な区別があるわけでもない。工学的な応用では，解析の中心は代数方程式や微分方程式，統計モデルなどによるモデル化と予測やシミュレーション，最適化であり，最終的な意思決定者が人間であることから，結果を人間に提示するための可視化も重要となる（図1）。

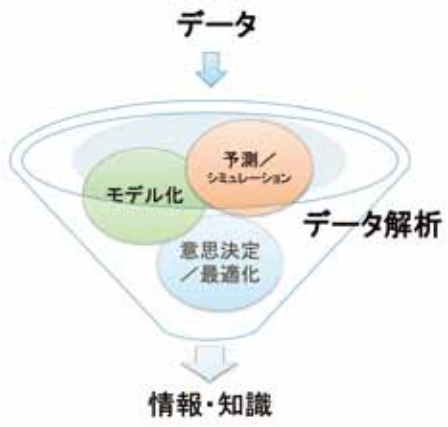

図1　データの取得から意思決定まで

データを解析する前には，当然，データの収集や蓄積，流通などの段階もあり，計測そのものが前提となることは言うまでもない。しかし，近年では，これらの段階は環境的に整っている場合が多くなっており，より解析そのものに集中できるようになってきている。従来，データ解析は，少ないデータから如何にして情報を取り出すかという問題に取り組んできたが，最近では，むしろ，データは氾濫し，情報過多の状態にある中，如何にして必要な情報を見つけ

Numerical Data Analysis and Process Systems Engineering
Yoshiyuki YAMASHITA（正会員）
1987年　東北大学大学院工学研究院博士後期課程修了
現　在　東京農工大学大学院工学研究院　教授
連絡先：〒184-8588　東京都小金井市中町2-24-16
E-mail　yama_pse@cc.tuat.ac.jp

2012年9月24日受理

† Miyanaga, K.　平成23, 24年度化工誌編集委員（12号特集主査）
東京工業大学大学院生命理工学研究科

るかという問題が重要になってきている。今日，データ解析のスペシャリストは，データサイエンティストとも呼ばれ，データを正しく評価して情報を引き出す者として広い分野において注目されている。

本稿では，数値データの解析手法の概要と，その果たしている役割について総括した後，化学プラントの運転・制御系におけるデータ解析手法の適用についていくつかの例を挙げて解説する。

2. 数値データの解析手法

データの解析は，文字や質的データのような非数値データの解析と数値データの解析に大別できるが，工学や自然科学で多用されるのは数値データの解析である。テキストマイニングも最近のトピックスではあるが，本稿では，紙面の都合もあるので，数値データの解析に限定して述べることとする。

実際のデータ解析にはソフトウェア（ツール）が不可欠である。簡単な解析であれば，表計算ソフトや汎用のプログラム言語で一から組み上げても良いが，専用のソフトウェアを使う方が格段に便利である。MATLABやSAS，SPSS，S-PLUSなどに代表される市販ソフトばかりではなく，ScilabやR，Wekaといったような高機能なフリーソフトが広く普及している[1-3]。

2.1 前処理

実世界のデータは，**表1**に示すようなさまざまな問題点を含んでいる[4]。その解析においては，前段階として，いわゆる前処理をおこなう必要がある場合がほとんどであるが，一般に，この処理はきわめて重要である。欠損値やはずれ値の処理や，ノイズの除去，変数の変換や次元の低減など，対象と目的に応じてさまざまな手法が適用される。たとえば，ローパスフィルタなどによる高周波ノイズの除去や，FFT（高速フーリエ変換）による時間領域から周波数領域への変換，PCA（主成分分析）やICA（独立成分分析）による変数の低次元化，kernel法による高次元空間への変換などは良く用いられる典型的な処理である。

表1 実世界データの問題点

多すぎるデータ	少なすぎるデータ	バラバラなデータ
・ノイズ	・欠損値	・データ間の非互換性
・はずれ値	・未測定変数	・複数のデータ源
・巨大なデータ量	・稀少な例	・精度の異なるデータ

2.2 回帰

データ解析の手法も実に多様であり，数千種類もの方法がハンドブックなどにもまとめられている。代表的なカテゴリーとしては，たとえば，回帰分析やクラス分類，クラスタリングなどがある[7]。どの手法を用いるかは，目的や対象に応じて適切に選択する必要がある。

回帰分析（Regression）は，最も古くからおこなわれている代表的なデータ解析の一つであり，従属変数（目的変数）と説明変数との関係を統計的な手法によって求めるものである。最も単純なのは線形回帰であり，最小二乗法によるパラメータ推定はあまりにも有名である。非線形関数を扱う非線形回帰や，複数の説明変数を扱う重回帰分析，SVR（Support Vector Regression）などさまざまな回帰手法がある。回帰モデルが求められれば，モデルに基づいて対象システムの解析やシミュレーション，最適化などが可能となる。

2.3 クラス分類およびクラスタリング

クラス分類（Classification）は，ラベル付きの訓練データを用いて，各ラベルを満たすパターンを導くものであり，決定木やルール導出，（教師付）ニューラルネットワーク，SVM（Support Vector Machine）などが代表的な手法である。学習後のモデルは，回帰モデルと同様に予測にも使えるが，クラス分類型の手法では出力が連続値ではなく，離散的なクラスとなる。

一方，クラスタリング（Clustering；Segmentation）は，与えられたデータに内在する特徴間の類似度に応じて，データを複数のカテゴリーに自動的に分類するものであり，SOM（自己組織化マップ）やART（Adaptive Resonance Theory）などの教師無しニューラルネットワークやk-means法などが代表的手法である。教師データにラベルがついている必要がない点がクラス分類型の手法との最大の違いである。

3. 化学プラントの運転・制御におけるデータ解析

3.1 プロセスモニタリング

プラントの運転・制御においては，数百から数千の測定値が毎分あるいは毎秒計測され，同時に数多くの操作変数を操作している。これらの膨大な運転データは，多くの有用な情報を含んでいるため，オンライン・リアルタイムでの活用だけではなく，オフラインでの活用のために，そのまま，あるいは圧縮されて記録されている場合も多い。これらの蓄積された運転データは，品質管理や生産管理，設備管理などの他，運転改善やプロセス改善などにも広く用いられている[5, 6]。

運転データ解析の典型的応用例である品質や運転状態の管理や異常診断は，時系列のトレンドデータを対象としたクラス分類型の問題である。品質管理手法としては，Shewhartチャート やCUSUM（Cumulative Sum），EWMA（Exponentially Weighted Moving Average）チャートに代表される統計的プロセス管理（SPC）が，古くから用いられている。こ

れらの手法においては，予め，既存のデータを解析してプロセスの平均や分散等を求めておき，現状が許容できるバラツキの範囲内に入っているかどうかを管理図で管理する。多変数の品質管理のためには，主成分分析などの多変量解析が用いられるようになっている。また，より高度な管理をめざして，判別分析やニューラルネットワークをはじめとするさまざまなデータ解析手法も適用されている。

3.2 プロセス制御

プロセス制御系の設計においては，モデル同定のためにデータ解析が多用されている。設定値や入力をステップ状に変化させ，その際の出力の応答を解析して制御パラメータやモデル係数などを同定する方法が最も一般的である。たとえば，モデル予測制御（MPC）では，予めステップテストなどによってシステムの応答を求めて，その結果をデータ解析して対象プロセスのモデルを作っておく。その上で，そのモデルを使って，将来のある時点での被制御量の値が目標値と等しくなるように，現時点での最適な操作量をオンラインで求めて制御をおこなう。

また，それぞれのプロセス制御系が，十分な性能で制御を実現できているかどうかを評価し，モニタリングするために，制御性能監視とよばれる各種のデータ解析がおこなわれている。制御系がオートモードとマニュアルモードに設定されている時間比を見るという単純な方法から，信号の変動パターンを解析してコントロールバルブに異常が無いかを判別する方法や，理想的な制御系の制御応答と比較する方法などが用いられている。

さらに，オンライン・リアルタイムに多変数の最適化計算をおこなって，PID制御系やMPCの設定値をダイナミックに変更するという，高度なデータ処理に基づく運転もおこなわれている。

3.3 ソフトセンサー

ソフトセンサーは，データ解析のオンラインでの典型的な適用例の一つである。訓練データに基づくデータ駆動型のモデルを用いた推定が，さまざまな対象に対して適用されており，プロセス制御やモニタリングに用いられている。化学プラントにおける産業応用例では，重回帰分析またはPLS（Partial Least Squares）といった線形の回帰モデルを用いたソフトセンサーが大半である。また，その主な対象プロセスは，蒸留プロセスと反応プロセスである[7]。最近では，特性変化や非線形性への対応のため，一部ではJIT型（Just-in-time）の手法も用いられるようになっている[8]。

4. おわりに

データ解析は，あらゆる計測データを正しく評価・解釈するために不可欠な解析である。データ解析技術は，分野横断的な基幹技術であり，これまで，さまざまな手法が，さまざまな分野で独自の発展を遂げ，大いに活用されている。

本稿では，数値データの解析手法を大まかに分類して，それぞれの特徴を簡単にまとめた後，化学プラントの運転・制御において用いられているデータ解析処理について，いくつかの特徴的な例を整理した。本特集号の他の記事と併せて，異なる分野において類似の解析手法が適用可能であることが見て取れるものと思う。各自が，それぞれの対象データを解析し，実データを正しく解釈する際に何らかのご参考となれば幸いである。

引用文献

1）橋本洋志：Scilabで学ぶ統計・スペクトル解析，オーム社（2008）
2）伏見正則，逆瀬川浩孝：Rで学ぶ統計解析，朝倉書店（2012）
3）Witten, I.H. *et al.*：Data Mining：Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann, Burlington, USA（2011）
4）Famili, F. *et al.*：*J. Intelligent data analysis*, **1**（1），3-23（1997）
5）山下善之 監修：計測・モニタリング技術—化学計測・計装の最先端とその応用，シーエムシー出版（2011）
6）山下善之：計測とモニタリング，化学工学，**74**（8），378-380（2010）
7）日本学術振興会プロセスシステム工学第143委員会ワークショップNo.27，高度プロセス制御に関するアンケート調査結果報告書（2009）
8）Kano, M. and K. Fujiwara：*J. Chem. Eng. Japan*, available online 15 September 2012, in press, doi：10.1252/jcej.12we167