品質・技術力向上に繋げる
SQC と機械学習のよりよい使い方について

トヨタ自動車㈱　業務品質改善部　渡邉克彦

１．はじめに

当社は約 70 年間、SQC(統計的品質管理)を問題解決の有効なツールとして位置づけ、ものづくりにおける品質・技術力向上に繋げてきた。近年、IoT の発展により機械学習が身近となり、SQCでは対応が困難であった問題の解決も可能となってきている。このこと自体は望ましいことであるが、機械学習はどのようなデータに対しても万能であると理解され、機械学習さえあればSQCは不要といった声も一部に聴かれる。今回、そのような現状に対し、目的や対象データに沿って両者のよりよい使い方を提案することで、ものづくりにおける更なる品質・技術力向上を目指していく。

２．ものづくりにおけるデータ分析（SQCと機械学習）の位置づけ

当部はTQMの推進部署として、研修展開・実践支援・啓発施策を通した品質・技術力向上の取組みを進めている。取組みの中で重視して伝えていることの1つに、問題解決の対象となる事象・システムにおいて、原理原則に基づいた仮説を構築し、実験等で得られた事実・データにより仮説が正しいかSQCで検証することがある。検証結果と仮説に差が認められれば、もう一度、原理原則から仮説を考え直し、再度検証を行っていく。この一連のプロセスが品質・技術力向上に繋がる「あるべきサイクル」となる。仮に原理原則に基づく仮説構築を行わず、SQCのみに頼ると誤った判断に繋がることも考えられる。例えば、少ないサンプルから得られた（見かけ上の）信頼度から目標を達成したと安心することや、最初から目標寿命を達成することだけを目指して寿命試験するなどである。これらは、目標を達成したとしても、その理由を説明できなければ、ノウハウや知見の蓄積などの品質・技術力向上には繋がりにくいともいえる[1]。近年身近となった機械学習においても、基本的にはSQCと同じデータ分析であることから、このサイクルは同じと考えており、得られた結果を原理原則で説明できる、一般解化できることがものづくりでは重要となる。

３．対象とする手法、データサイズ

さて、機械学習といっても現時点では様々な枠組みがあるため、今回は JUSE-StatWorks/V5 機械学習編に搭載されている手法を「機械学習」、JUSE-StatWorks/V5 総合編プレミアムに搭載されている手法を「SQC」として扱う（表1）。またデータサイズについても JUSE-StatWorks/V5 で扱える 1,000 変数×100,000 サンプル（ミドルデータの領域）までを対象とする。よって、それ以上のデータサイズ（ビックデータの領域）になると、それらの解析ツールとして著名な「R」や「Python」が必要となることを留意したい。なお、本論では JUSE-StatWorks/V5 に沿って、ものづくりにおける両者のよりよい使い方を提案するが、R や Python を使った場合でも基本的には同じことが言えるので参考にしてほしい。一般的に R や Python はデータ分析知識に加え、ある程度のプログラミング知識が必要となる。そのため、いざデータ解析に進むとプログラミングの壁にあたり、その解決策を探すのに時間を要するという苦労話をよく聴く。

トヨタ自動車　業務品質改善部　第 1TQM 室　グループ長
〒471-8571 愛知県豊田市トヨタ町 1 番地　TEL:0565-23-1048

表1． JUSE-StatWorks/V5 に搭載されたデータ分析の手法

解析の目的		SQC （総合編プレミアムなど 2011 年～）	機械学習 （機械学習編 2019 年～）
①データ可視化		多変量連関図、モニタリング	濃淡散布図、密度プロット、等高線図
②層別		階層的クラスター分析 非階層的クラスター分析(k-means 法)	混合ガウス分布
③情報の要約		主成分分析	カーネル主成分分析
④予測（回帰）		重回帰分析	正則化回帰分析（リッジ回帰、lasso 回帰、Elastic Net）
⑤分類	⑤-1	判別分析	サポートベクターマシン（SVM）
	⑤-2	AID、CAID	ランダムフォレスト
⑥外れ値検出		多変量管理図、MT 法	1 クラス SVM
⑦因果分析		グラフィカルモデリング	glasso

一方、JUSE-StatWorks/V5 は Excel に似たインターフェースに加え、プログラムを組むことなく手軽なマウス操作でデータ分析が可能であり、まさにパーソナルユースな操作性が強みである。PC 上で手軽に解析できること、これはデータ分析の普及・拡大には必要なことだと筆者は考えている。事実、当社でもSQCの普及・拡大を目指した 1980 年代に全く同じようなプログラミングの壁に当たったことからパーソナルユースな統計ソフト TPOS(Toyota Promotional Original SQC Soft）を開発し、社内普及に大きく貢献した[2]。ただし、いくらパーソナルユースとはいえ、やみくもに解析するのでは良い結果が出るとは限らない。SQCも機械学習も守るべき手順や、よりよい使い方があるので、次章より述べていく。

４．両者のよりよい使い方の提案

表 1 の解析の目的に沿って、ものづくりにおけるSQCと機械学習、両者のよりよい使い方について以下に提案する。

①データ可視化（SQC:多変量連関図など　機械学習：濃淡散布図など）

ここでは両者の比較というより、データ可視化の重要性をまず伝えたい。解析対象データを整理（欠損値を補完するなど）後、最初に実施すべきは基本統計量の確認であるが、その際、ヒストグラムや散布図でのデータ可視化を併用することで、基本統計量の数字を見るだけでは気付きにくい、外れ値（異常値）、層別の必要性、大まかな傾向等を確認しやすくなる。

ものづくりにおいて、これらはたいへん重要であり、データ全体を眺めることで、何らかのヒントを得るといっても過言ではない。ビックデータになって全体を見切れないから、いきなり解析に入っていくという話も聴かれるが、重要な情報を見逃すなどリスクも大きい。今回対象のミドルデータの領域であれば、必ず、基本統計量の確認とデータ可視化（以下基本動作と述べる）で全体に眼を通すべきある。例えば、相関係数が高い組合せ、それだけに注目していくことで予測したいという目的を達成することもありえる。筆者は 10 年以上、SQCを活用した問題解決の実践支援を社内・社外で実施しているが、高度な手法でいきなり解析したが、結果的に基本的な手法で事足りることも度々あり、これは機械学習についても同じと考える。

さて、今回、本論を執筆するにあたり、トヨタ・トヨタグループのTQM推進に携わる有識者にアドバイスを

頂きながら進めているが、これら基本動作の重要性が改めて認識できる話があるので紹介したい。トヨタ・トヨタグループでは昔から重回帰分析でできることは「予測」と「要因解析」と説明している。後者は各説明変数の標準偏回帰係数で目的変数への影響度が比較できるとしている。ここで、データ分析に詳しい方なら「要因解析」はできないと反対される方もいると思う。ではなぜ「要因解析」ができるとしているか。それはまさに、前述の基本動作を守ってきたからである。つまり、原理原則、固有技術で考えながらデータを取得する、基本統計量の平均値やばらつき(分散や標準偏差)が肌感覚と合うか確認する、外れ値、異常値があれば除外してよいか原因を確認してから外す、2つの集団が混在していれば層別する、相関係数の高いものは片方を除外する(多重共線性回避のため)などである。これらの手順を踏むことで必然的に解析にかけるデータの素性が良くなり、説明変数同士が直交に近い状態になるなど、要因解析を実施しても大きな問題とならないという理論である。

さらに変数選択においても、変数を1つひとつ重回帰式に取り込みながら技術的に考察する、新しい変数を取り込む際に、すでに取り込まれた変数の係数の符号が変わるのであれば、その理由を考える。このように変数選択を進め、最終的に出来上がった重回帰式を改めて原理原則、固有技術で考察する。式が説明できれば要因解析に使い問題解決に至るということである。もちろん、その後、実験計画法を使って詳細な寄与率を調べることも必要に応じて実施すべきであるが、これがまさに「実務で活きるデータ分析の使い方」だといえる。このように重回帰分析でも要因解析ができるように手順を整備し、苦労して定着させてきた先人方には敬意を表したい。

さて、話を戻すが、表 1 ①データ可視化では、機械学習には濃淡散布図、密度プロット、等高線図が搭載されサンプル数が多い場合の可視化機能が充実した。よって、データ分布の傾向把握や注目すべき変数を見つけるなど前述の基本動作に積極的に活用するのがよい。ただし、質的因子が扱えないため、質的因子が含まれる場合はSQCの多変量連関図を併用するのが望ましい。さらにn数が少なければ検定(平均値の差の検定、等分散性の検定、カイ2乗検定)機能もあるため層別等に活用すればよい。

変数やデータ数が多いと、ヒストグラムや散布図をしっかり見ること自体が難しくなるのは確かである。しかし、データサイエンティストの中にはそのような状況下でも時間をかけて1つひとつデータを可視化し、ヒストグラムや散布図を吟味する方もいるようである。筆者も大いに賛成であり、やはり、データが多いからといって基本動作を省略する、おろそかにすることは避けたい。解析のベースとなるので両者を使って、確実に基本動作を実施することが大切である。

②層別 (SQC:階層的クラスター分析など　機械学習 : 混合ガウス分布)

機械学習には混合ガウス分布、SQCには階層的クラスター分析、非階層的クラスター分析が搭載されている。混合ガウス分布は複数の多次元正規分布が重なったとして層別する手法であり、階層的クラスター分析や非階層的クラスター分析はサンプル間の距離等を使って層別していく。両者を使い分ける決定的な違いは、同じ座標での分布が重なるか重ならないかである(図1)。重なるのが混合ガウス分布で、重ならな

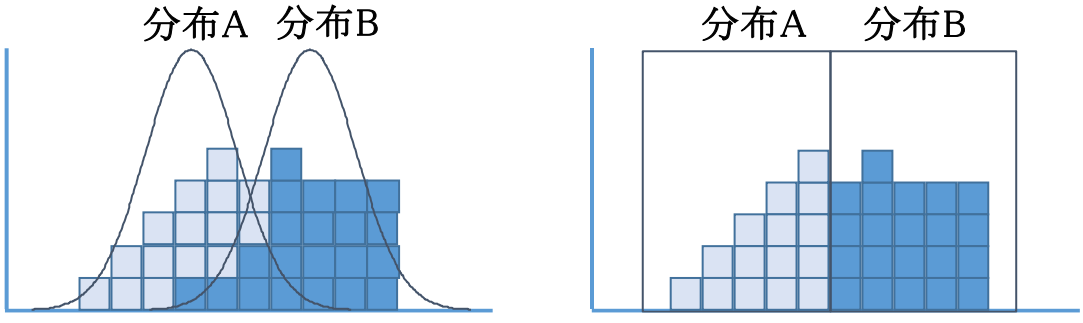

図1.　層別の違い(左:重なる　右:重ならない)

いのが階層的クラスター分析、非階層的クラスター分析となる。この重なる・重ならないは当然データ分析で決まるものではなく、原理原則、固有技術で決まるものである。つまり、原理原則、固有技術で考えて両者を選択すればよく、どうしても不明であれば併用し、結果を技術的に考察するのが望ましい。

③情報の集約（SQC：主成分分析　機械学習：カーネル主成分分析）

機械学習にはカーネル主成分分析が搭載され、SQCの主成分分析よりもサンプルを良く層別する場合があると JUSE-StatWorks/V5 のマニュアルにも謳っている。この説明だけではカーネル主成分分析が主成分分析よりも優れているように捉えられるが、解析の順番としては、最初に主成分分析であろう。主成分分析はビックデータに対しても解析に必要な計算（分散共分散行列の固有値問題）を問題なく解けることから、まずはそれで解析し情報要約・グルーピングに活用すればよい。

その後、カーネル主成分分析で、例えば主成分分析では見つからなかった異常値（高次元空間上ではじめて見つかる）の発見等に使えばよい。なお、図 2 に示す JUSE-StatWorks/V5 のサンプルデータ ML_M03_03 のようなドーナツ型の分布には主成分分析は歯が立たないとよく説明される（ここではカーネルパラメータ σ=0.49 が最も良く層別できるパラメータと決定している）書籍を見ることがあるが、①で述べたように、そもそもドーナツ型なのかどうかは基本統計量、多変量連関図のところで発見できるはずなので、その時点で層別等を実施すればよい。

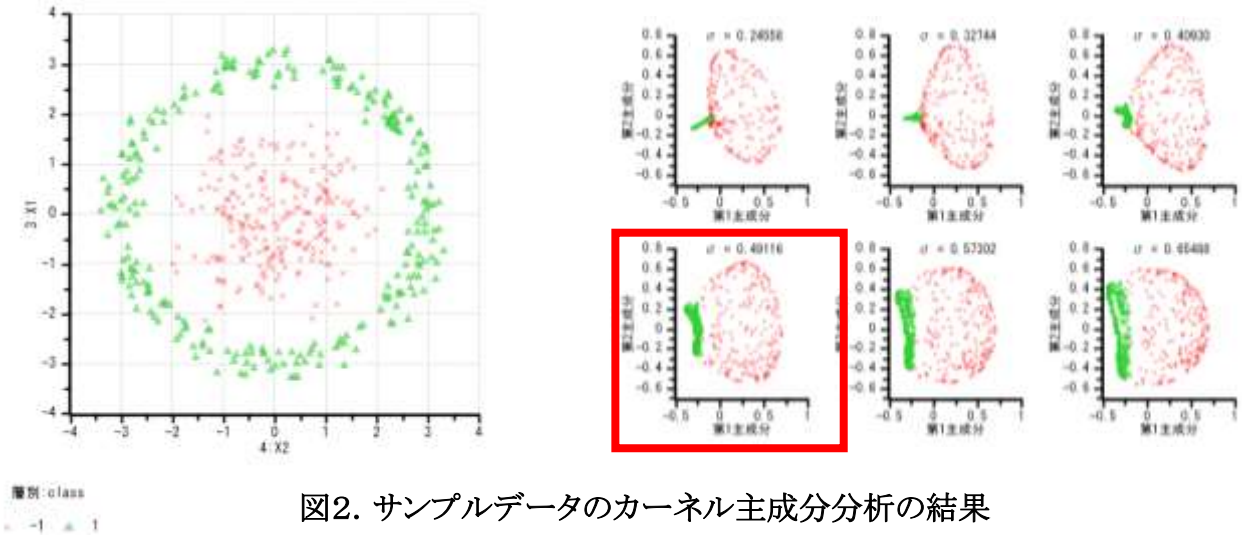

図2. サンプルデータのカーネル主成分分析の結果

④予測（SQC：重回帰分析　機械学習：正則化回帰分析）

回帰分析で予測をする際、単回帰分析の存在は無視できない。できる限りシンプルな式が望ましいというオッカムの剃刀という考え方があるように、単回帰分析で十分表現できかつ、原理原則、固有技術で説明できるのであれば、それを使うに越したことはない。単回帰分析では説明力が低いというのであれば、SQCの重回帰分析となる。ここでも原理原則、固有技術を意識しながら、変数を 1 つひとつ選択し、重回帰式を作り上げていく。必要であれば 2 次項や交互作用項も考慮して、自由度調整済み寄与率を確認しながら、納得できる重回帰式を作り上げればよい。ちなみに重回帰分析には予測誤差（将来データに対する誤差）を求められる機能も有しているので、将来データに対してどれくらい信用できるかの確認に使うべきである。なお、これらは前提としてデータ数 n が説明変数の数 p よりも大きい(トヨタでは n≧p'+20 を推奨。p'はモデルに取り込まれた変数)という

前提があることに注意したい。最近では IoT の発展によりデータが膨大にとれるようになり、p が n より大きいことも度々見受けられる。例えば p=10000 に対して n=100 のような場合では、重回帰分析では難しいとされている[3]。確かに、変数の増加に伴い、変数間に予想していなかった線形制約（例えばシンプルな例だと X1+X3+X5=X7 など）によって回帰係数を求めるのに必要な逆行列を求められないことも十分に考えられる。そこで、機械学習の正則化回帰分析（リッジ回帰、lasso 回帰、Elastic Net）が登場である。大切なのは、特に p>n などの状況になっていないのに、まず真っ先に正則化回帰分析は順番が違うということを伝えたい。なお、正則化回帰分析は確かに強力な手法であるが、荒木ら[3]が述べるように、取り込まれた回帰係数は正則化によって「縮小」されていることがあるので、その解釈には注意したい。また、変数選択においても重回帰分析の変数選択のように意のままに取り込まれる変数をコントロールできないことも念頭に入れたい。

なお、機械学習には「Leave-one-out 法、k-分割交差検証法、ホールドアウト法」が搭載されている。ビックデータにおいては過学習が避けられないため、これらでモデルの汎化能力（モデルの良し悪しや予測精度）を評価することが必要である。

⑤-1 分類（SQC：判別分析　機械学習：サポートベクターマシン）

どちらの群に属するかの分類に対して、機械学習はサポートベクターマシン（以下 SVM）、SQCは判別分析が搭載されている。両者の違いとして、判別分析は全データを用いて各々の群が正規分布を前提として判別境界を作成することに対し、SVM は境界付近の点（サポートベクター）だけを使って判別境界を作成する。これに加え、SVM はカーネルトリックを使うことから強力な判別力があることは間違いない。ただし、図 3 に示すように、SVM（ガウスカーネルを採用）は正則化パラメータ C の値によっては、針の穴を通すような判別境界も引けてしまう(図 3 の C=1000)。このような過学習を心配しなければならない判別境界となるなど、パラメータ決定は難しいといえる。一方、判別分析は判別式が求められる（図 3 の判別分析結果の判別境界線。判別式 Z=0.55*強度縦-1.74*強度横+66.98）ため、説明変数が 1 単位増えるとどちらの群に向かっているか式で解釈できるなどのメリットもある。

まとめると、ものづくりによくある各々の群が正規分布に近ければ判別分析がよい。SVM がいくら強力な判別力があろうとも、判別分析で事足りるのであればわざわざ SVM を使う必要はない。判別分析で解析した結果、判別力が不十分ならば SVM で解析すればよいが、判別式はないため、判別境界の解釈を原理原則、固有技術で説明できることが必要となる。

⑤-2 分類（SQC：AID、CAID　機械学習：ランダムフォレスト）

どういう条件で分岐（層別）されるかの分類には、機械学習はランダムフォレスト、SQCは AID, CAID が搭載されている。AID、CAID は、連続した層別が強力な武器であり、分岐が可視化されるため、解析者が原理原則、固有技術で解釈しやすいというメリットがある。一方、ランダムフォレストには AID，CAID とほぼ等価な機能を有する決定木が解析プロセスに含まれ、AID、CAID と同じような分岐の可視化が可能となる。そのため、ランダムフォレストを使えば事足りることになる。ちなみに、各手法は分類基準が異なり、AID は F 値、CAID はチュプロウの t 値、ランダムフォレストの決定木はジニ係数を用いている。なお、AID、ランダムフォレストの決定木は分岐数が 2 で固定されているが、CAID だけは 3 分岐以上に対応しているため、解析上都合がよい場合は CAID を使うのがよい。

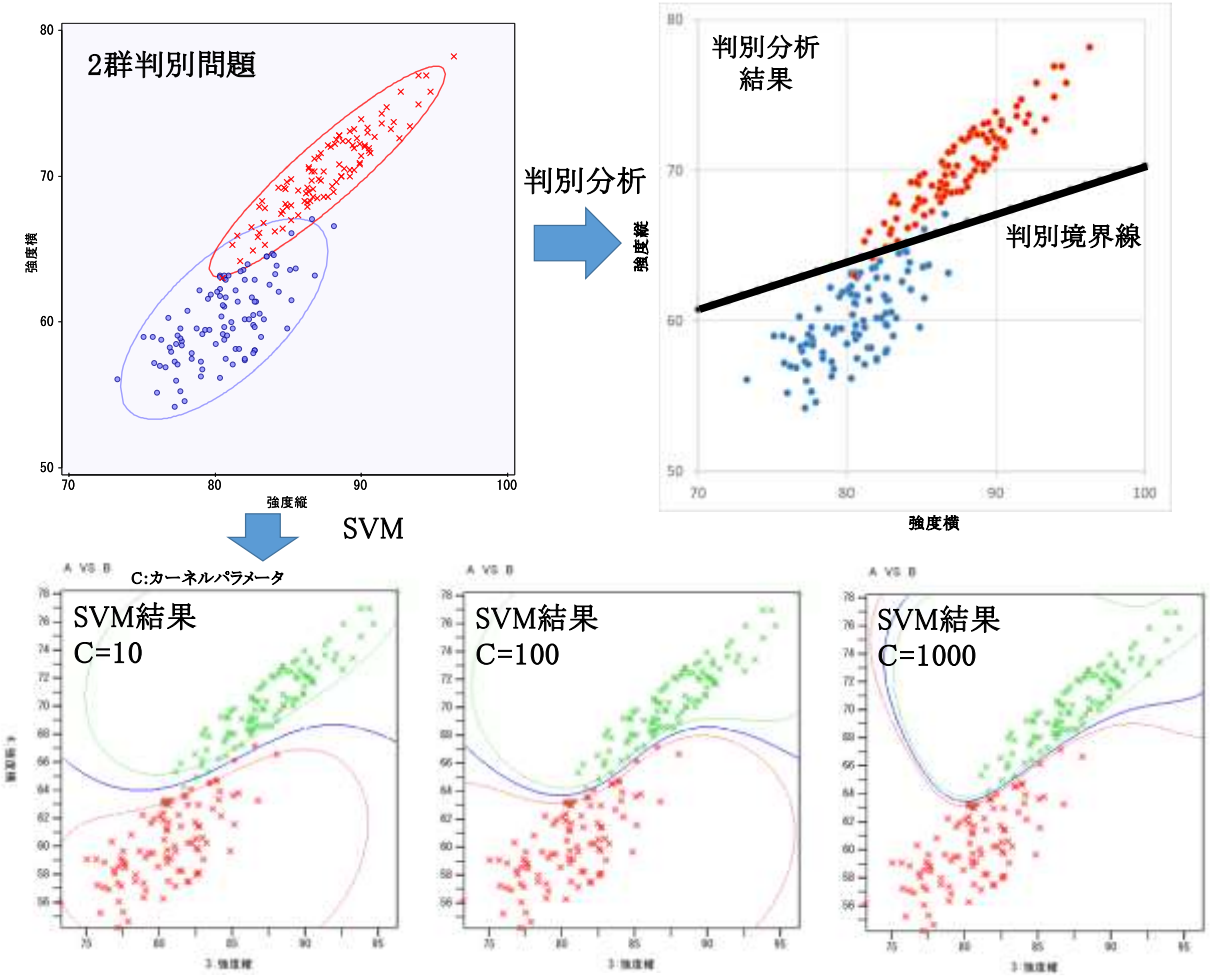

図 3.　2 群の判別問題に対する、判別分析と SVM の比較

⑥外れ値検出（SQC：多変量管理図、MT 法　機械学習：１クラス SVM）

ここでの外れ値検出の定義は、前処理における外れ値削除よりは管理図での異常（いつもと違う状態）を発見することとする。機械学習は 1 クラス SVM、SQCは多変量管理図（MT 法）であり、両者の比較には JUSE-StatWorks/V5 のサンプルデータ M5_6_1 を使用する。100 日分の管理特性 1 と管理特性 2 のデータで、標準空間が図 4 内記載のように設定されている。まず、多変量管理図の結果を図 4（左）に示す。管理限界線であるカイ 2 乗分布の上側 0.27%点(1 変数の 3σ に相当)を楕円で示しており、これより外側にあるデータ No98～No100 が異常として検出されている。つまり、マハラノビス距離をみて異常と判断していることになる。一方、図 4（右）に 1 クラス SVM でのほぼ同じ検出条件（偽陽性率 3%、カーネルパラメータ C=3.5）の結果を示す。丸で示す 3 点が異常として検出されたが、多変量管理図と全く違う点（No34,65,91）となっている。カーネルトリックで高次元空間に写像することでこれらの点が異常と判定されたのだが、技術的に解釈することはかなり難しいと考える。詳細は割愛するが、カーネルパラメータ C の値を変更することで検出される点は変化することもあるなど使い方には細心の注意が必要である。

まとめると、ものづくりでよくある正規分布に対して外れ値検出を行うのであれば、まずは多変量管理図や MT 法（原理としては両者は同じ）を実施し、マハラノビス距離で外れ値（異常値）を発見すればよい。一方、1 クラス SVM はパラメータ C の値（ガウスカーネルの場合）によって検

出される点も異なる上、そもそも標準空間、分布という概念が無く、与えられたデータの外側α％（αは任意）は外れ値という考え方のため、そのような状況下にて使う手法となる。

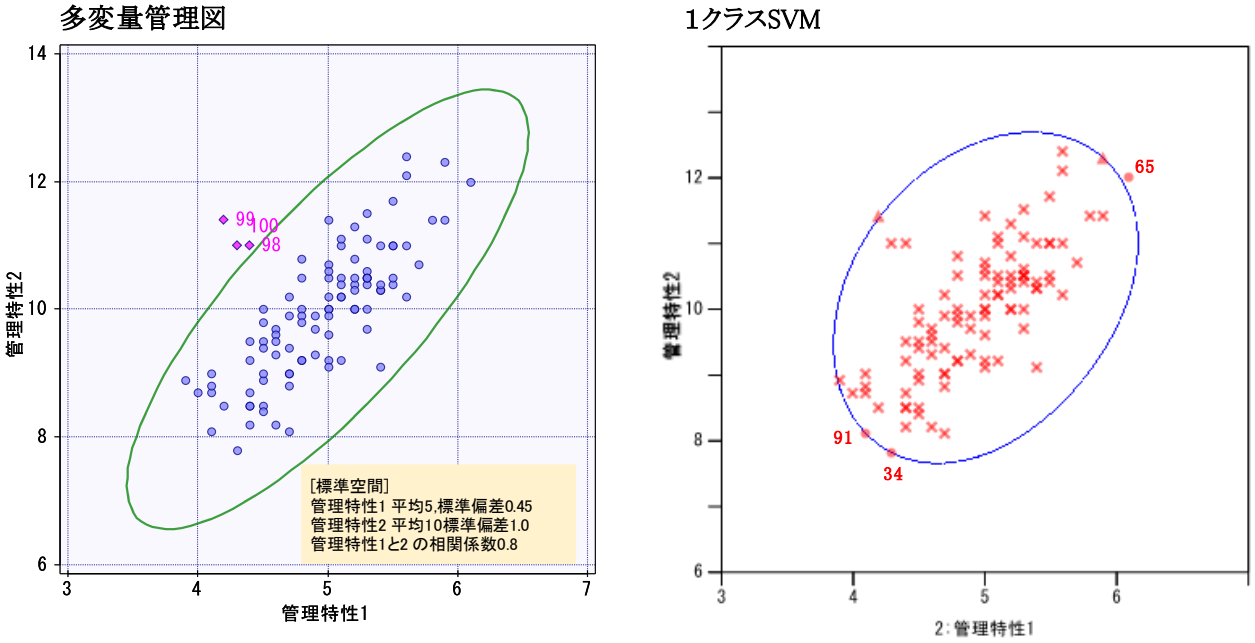

図 4. 多変量管理図での管理限界線（上側 0.27％点）を越えた点（左）
と 1 クラス SVM での偽陽性率 3%、C=3.5 で検出された点（右）

⑦因果分析（SQC：グラフィカルモデリング　機械学習：glasso）

グラフィカルモデリングに正則化を加えたものが glasso であるが、④で述べたようにいきなり、機械学習の glasso で解析するよりは、まずSQCのグラフィカルモデリングで解析するのがよい。さらに好ましいのは、JUSE-StatWorks/V5 には基本解析ーモニタリングに相関係数行列を可視化する機能もあるため、最初に変数間の相関関係について把握しておくとよい。つまり、まず相関係数行列で全体を把握し（偽相関も含まれることを注意しながら）、次にグラフィカルモデリングにて偏相関の関係を把握する。ここで、最初の相関係数行列で相関係数の高い組合せが多く存在するのなら、グラフィカルモデリングでの偏相関係数も安定しない可能性があるため、更に glasso で解析し、グラフィカルモデリングと同じような変数が取り込まれるのか確認していけばよい。これにより変数間の詳細な関係性を確認することができる。

また、グラフィカルモデリングや glasso では目的変数に強い偏相関があり、かつ他の変数とは弱い偏相関である説明変数を発見することができる。よって予測が目的であれば、そのような説明変数を抽出して重回帰分析や正則化回帰分析で予測式を作成すると、他の変数に影響されにくいロバストな予測式となるので推奨したい。

５．まとめ

今回提案した、ものづくりにおけるSQCと機械学習のよりよい使い方を表 2 にまとめる。両者は同じデータ分析のツールなのだから、片方だけを使うのではなく、適した場面で適した手法を選択することで、ものづくりにおける品質・技術力の向上に繋がっていく。

表 2.　JUSE-StatWorks/V5 におけるSQCと機械学習のよりよい使い方　まとめ

解析の目的		SQC （A）	機械学習 （B）	本論で提案した よりよい使い方
①データ可視化		多変量連関図、 モニタリング	濃淡散布図、 密度プロット、等高線図	基本は(B)。質的変数があれば（A）も併用する
②層別		階層的クラスター分析 非階層的クラスター分析	混合ガウス分布	同一座標で重なる(B)、 重ならない(A)で選択
③情報の要約		主成分分析	カーネル主成分分析	基本は(A)→(B)
④予測（回帰）		重回帰分析	正則化回帰分析 (リッジ回帰、 lasso 回帰、Elastic Net)	基本は(A)→(B)
⑤分類	⑤-1	判別分析	サポートベクターマシン（SVM）	群が正規分布とみなせるなら(A)、それで不十分なら(B)
	⑤-2	AID、CAID	ランダムフォレスト	(B) で十分だが、(A)の CAID は 3 分岐以上に対応
⑥外れ値検出		多変量管理図、MT 法	1 クラス SVM	正規分布ならば(A)が解釈しやすい。それで不十分なら(B)
⑦因果分析		グラフィカルモデリング	glasso	基本は(A)→(B)

謝辞

本執筆を進めるにあたり、有益なアドバイス・ご意見を賜りました日本科学技術研修所、トヨタグループデータ分析分科会の皆様には謝意を表します。

特に㈱日本科学技術研修所　犬伏秀生氏、㈱豊田自動織機　久保田享氏、松山立氏、トヨタ自動車九州㈱　佐々木康博氏、則尾新一氏、トヨタ自動車㈱　小杉敬彦氏、阿部誠氏、田中公明氏、佃茂昭氏には草稿の段階より丁寧なアドバイスを頂きました。厚く御礼申し上げます。

参考文献

[1] 渡邉：品質・技術力向上にむけた信頼性データ解析の応用、日科技研、第 2 回信頼性データ解析シンポジウム、P7～13、2012 年

[2] 天坂他：仕事の流れにとけこむＳＱＣソフトを目指して、日本品質管理学会　第 25 回年次大会発表要旨集、P3～6、1995 年

[3] 永田、荒木他：日本品質管理学会中部支部　産学連携研究会編　開発・設計に必要な統計的品質管理、日本規格協会、P191～206、2015 年

■公開資料ページ
弊社ウェブページで各種資料をご覧いただけます http://www.i-juse.co.jp/statistics/jirei/

■お問い合わせ先
(株)日科技研 数理事業部 パッケージサポート係 http:/www.i-juse.co.jp/statistics/support/contact.html