アナリティクスにおける探索的データ解析

末次浩詩／平山文洋

データ分析ツールを活用する企業が増えている一方で、十分にデータを活用できていないという声が少なくない。その理由の一つに、個別事情を多く含む企業の業務データは、従来の統計解析的なアプローチだけでは十分な分析ができないことが挙げられる。そのような際は、さまざまな手法でデータを可視化し、その特徴を広く捉える探索的データ解析が有効で、現場が納得できアクションにつながる分析結果が得られる。データ活用においては「まずやってみる」ことが重要だが、それが実施しやすい点でも有益なアプローチである。

データ分析における「解釈」の難しさ

近年、データ活用を経営課題と捉える企業が増える中で、データ処理のためのIT基盤や分析ツールの導入が進んでいる。しかし、大量データを高速に処理する基盤と、高度な統計計算を行う分析ツールを揃えただけでは、データ活用は進まない。データをもとに最終的に意思決定するのは人であり、意思決定するには、処理されたデータを「解釈」しなければならないからである。

明確な仮説設定さえあれば、統計的な検定にかけることで、有意水準などの定量的な確率をもって結果に対する判断が可能である。ところが、この仮説設定が適切にできなければ、算出されたデータを解釈することができず、せっかくの解析ツールが意思決定につながらず、データ活用が進まないのである。

これは、ビジネスの構造が複雑で、データが現場や顧客の個別事情を多く含んでいることに起因している。たとえば、小売業や製造業のプロモーションの効果を測定しようと思えば、「いつもよりどれだけ売れたのか」を評価するだろう。しかし「いつも」をどう定義すればよいのだろうか。厳密には、「プロモーションを実施したかしなかったか」を除く条件がすべて同一の「いつも」は存在しない。こういうケースでほとんどの企業が採用している評価指標は昨対比（対前年比）である。しかし、実際は市場に投入する商品も商圏も競合環境も前年とは変化しており、昨対比で評価しても根拠が十分でなく、現場が納得できるアクションにはつながりにくい。

データ解析の出発点である探索的データ解析

一般的に、データ解析には２つのアプローチがある。「確証的データ解析」と「探索的データ解析」である。確証的データ解析は仮説検定のアプローチであり、ある仮説が正しいかどうかを統計的・確率論的に検証するものである。探索的データ解析は、1960年ごろに米国の統計学者J.W.Tukey（テューキー）によって提唱されたもので、「データを視覚的に捉え、データ自身に仮説を語らせるアプローチ」と言える。確証的データ解析が前提とする「仮説」を立てるうえで、データの構造・分布をグラフを通じて正確に捉えようとするものである。

データ解析において明確な仮説設定ができていない状況では、まず探索的データ解析によってデータの特徴を捉えることが有効である。これは、市場環境が激しく変化している場合や、センサーデー

当レポートに掲載されているあらゆる内容の無断転載・複製を禁じます。すべての内容は日本の著作権法および国際条約により保護されています。
Copyright©2014 Nomura Research Institute, Ltd. All rights reserved. No reproduction or republication without written permission.

タのような、これまで扱ってこなかったデータを対象とするような場合には特に有益な方法である。

Tukeyが提唱した「箱ひげ図」（長方形とその両側に伸びた線を用いて、最小値、第１四分位点、中央値、第３四分位点、最大値を表した図）は１変量のデータ分布（データの全体的な特徴）を捉えるのに役立つ。ヒストグラムが１つの属性値についてのデータ分布を表現するのに対し、箱ひげ図は属性の異なる複数のデータの分布を表現できるため、１つのグラフで効率よくデータの特徴を捉えることができる。

２つの変量の関係を正確に捉えるためには「散布図」（横軸と縦軸にそれぞれ別の量を取り、データが当てはまる位置に点を打ったグラフ）が、変量が３つある場合は「バブルチャート」（散布図の横軸、縦軸の量に加えて、もう１つの量を円の大きさで表して配置したグラフ）が有効である。

ウェブベースのダイナミックな可視化ツール

グラフを用いて探索的データ解析を実施していると、「外れ値」（ほかの多くとかけ離れた値）や「異常値」（あり得ない値）に出くわすことがある。たとえばサプライチェーンの在庫分析であれば、外れ値や異常値を示すデータがどの商品なのか、どの拠点なのか、といったことを特定したくなる。このような時、可視化を通じた探索的データ解析がインタラクティブにできるツールがあると非常に便利である。以下では、野村総合研究所（NRI）における、可視化ツールを利用したデータ活用の検討事例を３つ紹介したい。

(1) サプライチェーンにおける在庫分析

在庫が適正かどうかを評価するうえでは、欠品率と在庫水準が主要な指標となる。しかし、商品の特性や仕入れ条件などから在庫を多くせざるをえなかったり、販促活動の結果として欠品が生じたりすることがあり、見えている指標だけでは正しく評価できないことが多い。

そのため、在庫分析を実施する際には、拠点別や商品別に、欠品率や在庫水準などのサプライチェーン上のデータの全体観をつかんだうえで、各拠点や商品ごとのそれら指標の時系列推移、さらには単日の受注明細までをインタラクティブに検証できる環境を持つことが必要である。

次ページの図１は、NRIの可視化ツールの画面の例で、拠点別・商品別の指標数値を表す散布図から、時系列推移、単日の受注明細といった具合にデータを詳細化して表示したものである。販促活動の実績も、実施期間に応じた長さのバー（中段の図の上部にある棒状の図形）で表現している。そこにカーソルを当てることにより、どのようなキャンペーンをいつからいつまで実施したのか、といった定性情報を確認することもできるようになっている。

(2) 顧客動向分析

「ID-POS」に代表される、顧客を識別できる購買データが取得できるようになったことで、企業は顧客ごとの来店頻度や購入金額などを正確に把握できるようになり、顧客にステータスを付与してそれに応じた施策を実施できるようになった。

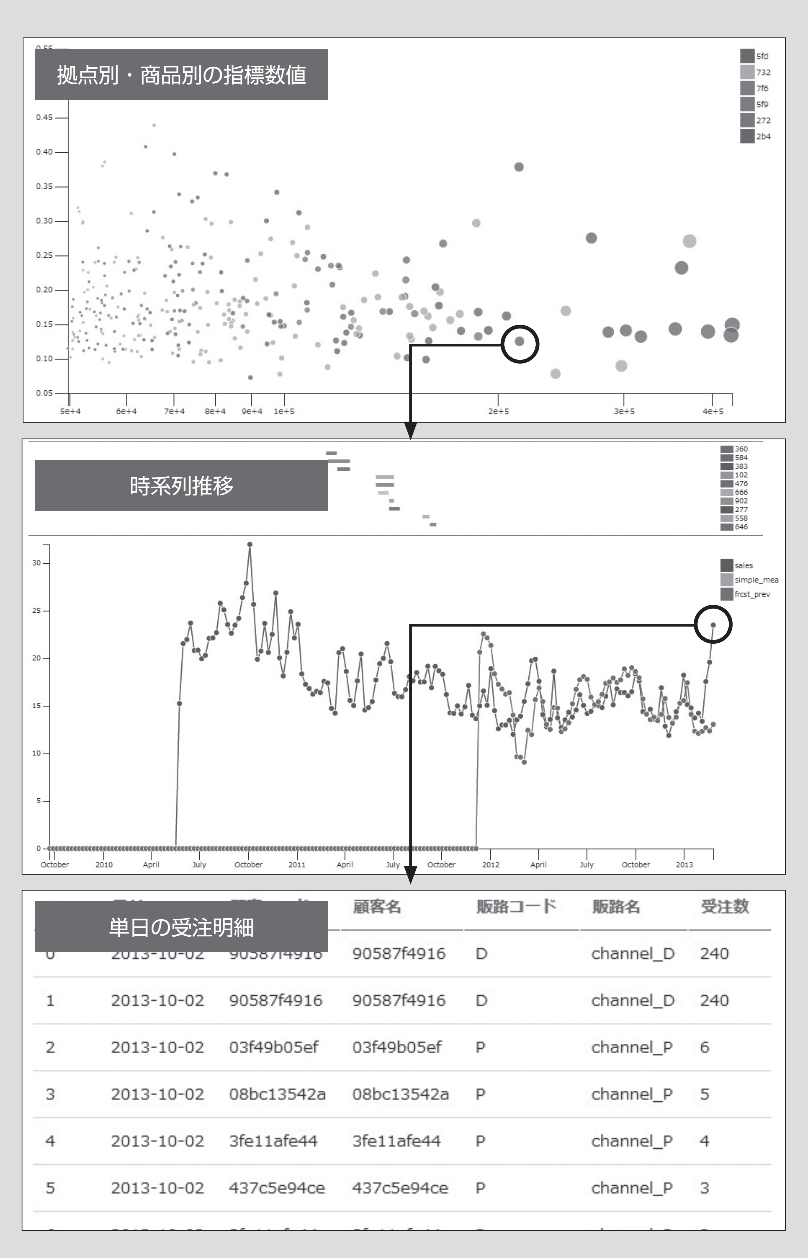
企業が実施するさまざまなキャンペーンやその他の施策に効果があるのかどうかを検証するためには、ある時点であるステータスにいた顧客が、翌月・翌四半期・翌年にどのステータスになっているか、すなわち「個」客の「フロー」

当レポートに掲載されているあらゆる内容の無断転載・複製を禁じます。すべての内容は日本の著作権法および国際条約により保護されています。
Copyright©2014 Nomura Research Institute, Ltd. All rights reserved. No reproduction or republication without written permission.

を追うことが求められるが、フローまで追えているケースは少なく、ステータスの構成、すなわち「ストック」の推移を確認するにとどまるケースが多い。

しかし最近は「Sankey（サンキー）ダイヤグラム」と呼ばれるチャートが登場し、ストックとフローの情報を同時に表現することが可能である。図2は、可視化ツールを使って顧客ステータスの変化をSankeyダイヤグラムで表したものである。新規入会した顧客群を表すバー（上段のグラフの左上の棒状図形。バーの長さが顧客数を表す）をクリックすると、それらの顧客がその後どのようなステータスになっていくかを、人数とともに示した詳細な図を表示させることができる。

図1　野村総合研究所の可視化ツールによる在庫分析の例



(3) 多変量構造の把握

多変量のデータは、その構造を正確に把握することが極めて難しい。因果と相関の区別は困難であるし、実際には相関関係がない2変量について散布図を描いてみると、あたかも相関があるかのように見えてしまう、いわゆる「見せかけの相関」（共通の決定要因が存在すること）にも注意する必要がある。

多変量解析では、変量間の関係を表す指標に偏相関係数がある。これは、２つの変量の関係について、他の変量の影響を排除した相関を示す指標である。偏相関係数は母集団としてある程度のサンプ

当レポートに掲載されているあらゆる内容の無断転載・複製を禁じます。すべての内容は日本の著作権法および国際条約により保護されています。
Copyright©2014 Nomura Research Institute, Ltd. All rights reserved. No reproduction or republication without written permission.

ル数が必要なうえ、外れ値の影響を大きく受けるため、突発的な出来事によって相関が過剰に高く算出されることも多い。

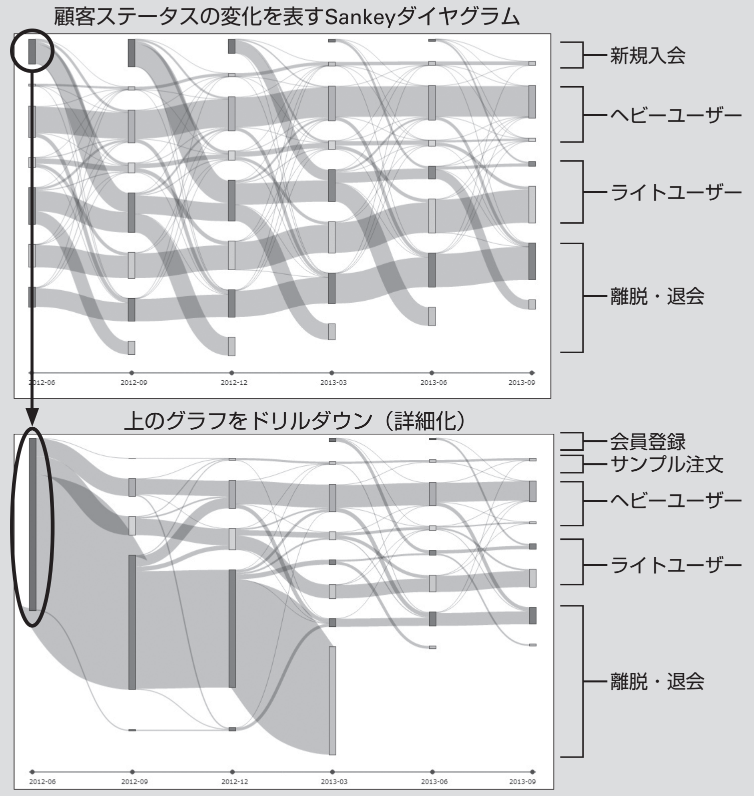
このような多変量解析においても可視化は有効である。誌面の都合で図は示せないが、最近の可視化技術で「平行座標プロット(parallel coordinate)」という図表を表現すると、1つの変数の数値を動かしていくと他の変数がどのように変化するかを、視覚的に検証することができる。

スモールスタートのデータ活用

データ活用を進めるうえで明確な仮説設定ができないということは、データ活用の効果が事前に推定できないということである。従来であれば、このような状況で大規模な投資を行うことは難しかった。しかし最近では、クラウドサービスの拡大によって、大量データを高速に処理するための環境を比較的低コストで構築することも可能となった。

もちろん仮説設定が大事であることに変わりはないが、データの構造も分布もわからない状態で抽象的な構想ばかりを進めていては、現実的な収益化の方向性は出てこない。データの活用においてはスモールスタートの姿勢、すなわち「まずやってみる」ことが重要である。仮説設定や収益化の方向性についても、実際のデータを眺めながら継続的な検討を行い、場合によっては初期仮説にこだわらずに方針を転換するという柔軟さが必要ではないだろうか。本稿で紹介したダイナミックな可視化技術を活用した探索的データ解析は、「まずやってみる」ための有効なアプローチになるはずである。

図2　可視化ツールによる顧客動向分析の例



『ITソリューションフロンティア』2014年８月号より転載

末次浩詩（すえつぐひろし）
産業ビジネスデザイン部主任コンサルタント

平山文洋（ひらやまふみひろ）
産業ビジネスデザイン部上級コンサルタント

当レポートに掲載されているあらゆる内容の無断転載・複製を禁じます。すべての内容は日本の著作権法および国際条約により保護されています。
Copyright©2014 Nomura Research Institute, Ltd. All rights reserved. No reproduction or republication without written permission.