ビックデータと統計学

中央大学 理工学部 教授　鎌倉　稔成

昨今、ビッグデータという言葉を多く耳にする。新聞の一面にはビックデータを利用した優良顧客の抽出のビジネスモデルというような記事が何度となく目にする。

果たしてビジネスにつながるような、ビッグデータの活用によるうまい話は可能なものだろうか。大規模なデータから意味のある情報を抽出する技術を知らなければ、データはゴミと同じである。「ゴミの山から金を」というようなキャッチフレーズによるデータマイニングも十分にビジネスの場で活用されただろうか。本稿では、ビッグデータ活用のために統計技法がいかに必要であるかについて解説する。

ビッグデータとは

2000年代の初頭、データマイニング[1]という言葉が流行した。統計学が扱うデータの量は急速に肥大化してきた。これは、コンピューター技術の進歩に伴うものである。Huber[2]によると、データのサイズの分類は**表1**のようになっている。

表1　データの大きさと記憶媒体

分類	サイズ	記憶媒体
Tiny	10^2 bytes	紙
Small	10^4 bytes	数枚の紙
Medium	10^6 bytes	フロッピーディスク
Large	10^8 bytes	ハードディスク
Huge	10^10 bytes	ハードディスク
Monster	10^12 bytes以上	ハードディスク、テープ

筆者はデータマイニングの扱うデータとビックデータの扱うデータは本質的には同質と考えているが、ビッグデータという言葉が現れるまでは、データマイニングの台頭からほぼ10年と考えられるので、その10年の間のコンピューターのデバイス技術が反映されていると考えて良い。水田[3]によると、ビッグデータが登場したのは2007年頃からであり、注目を集め始めたのは2011年のマッキンゼーレポートからとしている。Google*1 insights for search のグラフでは2010年頃から急速にその人気度を上げてきている。特に、インターネットの普及はめざましいものがあり、インターネット上のデータの蓄積を１つの大きな違いとみることができる。Twitter*2 やFacebook*3 のデータは莫大なものとなっている。

データを量だけで分類するのであれば、数テラバイト（TB）を越えるデータをビッグデータと考えてもよい。しかしながら、これは、あくまでも、計算機環境との相対的なものであり、パソコンレベルでは数ギガバイト（GB）でもビッグデータといってもよい。計算機のハンドリングに時間がかかりやっかいなレベルと考えるのが良さそうである。もとより、統計学においては、最尤法の性質を調べるには、極限演算が主流である。データサイズで見れば、無限ということである。無限といえば、データのサイズが無限ということであり、ビッグデータ以上のものであるが、このように、理論統計が扱う、極限演算としての無限のデータはビッグデータとは言われない。例えば、コインを投げた時、表の出る確率を推定したいとする。10回投げて、4回が表、6回が裏、表の出る確率は4/6=0.6と推定されるが、投げる回数を100、10000、...と増やしていった時どうなるかということである。投げる回数をどんなに増やそうと（増やしていけば、数TBのデータとなることは容易に想像できる）、ビッグデータとは呼ばないということである。

大きさだけで、定義できない何かがあるのである。統計学では、母集団というものを定義する。母集団は、狭義には、あまり多くないパラメーターを含む確率モデルに寄って規定する。コイン投げの例では、

$$f(x) = p^x(1-p)^{1-x} \ (x = 0 \ or \ 1)$$

というベルヌーイ分布を仮定する。このモデルにおけるパラメーターはもちろん１つ、pである。観測値を$x_1,...,x_n$と表せば、pの推定値はs_nを総和としたとき、$\hat{p}_n = s_n / n$となる。初等数理統計学の知恵を借りれば、

＊1）「Google」は、Google Inc. の商標または登録商標です。　＊2）「Twitter」は、Twitter, Inc. の商標または登録商標です。　＊3）「Facebook」は、Facebook, inc. の登録商標です。

$\hat{p}_n$は不偏かつ一致推定量である。一致推定量は$\hat{p}_n$がpに確率収束することを意味する。つまり、莫大なデータを扱うといういい方もできる。しかし、これだけではビッグデータとはならない。統計学の立場でビッグデータを定義すれば、データが非常に複雑な機序のもとに生成されるようなデータということになる。コイン投げの例を用いると、コイン投げのときに共変量と呼ばれる補助データを観測する場合を考える。例えば、画像、温度、湿度、空気の粘性、指の位置、センサーによる指の加速度等のデータがあったらどうだろうか。このデータを100万回も取ったら、十分ビッグデータといえる。もはや単一のパラメーターpではモデルは記述できない。非常に複雑なものとなる。

統計学の立場でのビッグデータの定義は試みたが、世の中一般の定義としては、水田を引用すれば、3つのVで表されるというイメージが妥当である。3つのVとはVolume, Velocity, Variety である。Volumeはすでに述べたが、データの量（サイズ）である。Velocityはデータが高速に入手されなければならないことを示し、例えば、インターネットのアクセスデータやSNSデータのように時々刻々とデータが蓄積されることを示している。したがって、大容量、高速ストレージが必要であることは言うまでもない。Varietyはまさに、多様性のあるデータ、同質でないデータということになる。

ビッグデータに何が必要か

ビッグデータとしてどのようなものがとられているかを見てみよう。

- ウェブサイトリンク
- twitterの反応データ
- 画像データ
- テキストデータ
- e-mail
- 製品レビュー
- ムービーデータ

また、大きな特徴として、時系列になっているということである。さらに、ここに挙げられたデータ時系列上に複雑に絡み合っているという特徴もある。時系列上に頻繁にデータが観測され、それらが互いにコミュニケーションを取りながら相互作用を及ぼしている。インターネットの高速化がデータの生成過程において、Velocity をさらに高速化させている。

分析のためには、高速かつ大容量、安全なストレージが必要であり、データの高速書き込み、読み出しのアルゴリズムの研究も盛んである[4]。

ビッグデータの統計的扱いの難しさ

ビッグデータの扱いの難しさは、統計処理を行う以前の問題もたくさんある。データをどのように蓄積し、また、高速に取り出すかというデータベースの設計上の問題、さらに、データベースエンジンと統計処理プログラムとの接続上の問題等、統計解析を行う以前にすべきことがたくさんある。

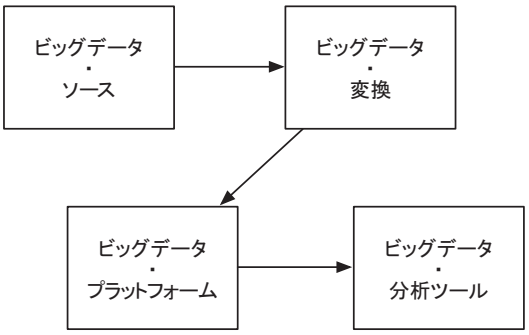

図1　ビッグデータの分析までのフロー

大きなフロー（**図 1**）は、ソースから粗データを変換（テーブル化）し、Hadoop[4] やMahout[4] のようでデータベースソフトに送り込み、そこから分析ツールを利用するというものである。ビッグデータの分析には分析するデータの抽出作業が不可欠で、かつ、高速に行う必要がある。

さて、分析を行うための、統計学本来の難しさは、データサイズが多いためのハンドリングもあるが、データの中の特定のIDの抜き出し等、高速に行うことが可能なものは、プラットフォームに任すことが必要である。検索などのデータ抽出を統計計算ソフトウェアで行うと莫大な時間を消費してしまうことが多い。それでもなお、比較的大きなデータをプラットフォームに頼らずに処理したい場合はassociative arrayやsparse arrayを用いてデータのインデクス化をしておくのがよい。いくつかの言語には実装されているので、必要に応じて使用してみてもよい。

さて、統計学の問題に戻ろう。統計学は本来データから効率的に情報を抽出するための学問であり、また、分析者や意思決定論者のサポートを行うことを目標としている。したがって、分析者のデータに対しての仮説表現を行うことが重要である。ビッグデータの難しさは、分析者は仮説表現としてのモデルすら、最初の局面では持てないことが多いということに起因する。必然的に、確率ベースの精密なモデリングの前に、データを探索的に眺め、全体を俯瞰するという作業が必要になってくる。

*4）「Hadoop」、「Mahout」は、Apache Software Foundation の米国およびその他の国における商標または登録商標です。

次にマイニング技術が必要になる。冒頭述べたが2000年初等に脚光を浴びた、統計的マイニング手法と呼ばれるいくつかの方法を利用する。

- マーケットバスケット分析
- 決定木（分類木）
- k-means クラスタリング
- RFM分析
- ロジスティック分析
- 重回帰分析
- ニューラルネットワーク
- マシンラーニング
- サポートベクターマシン
- テキストマイニング

これらの方法はデータマイニングとして脚光を浴びた方法であるが、もちろん、2000年頃に開発されたものではなく、長い歴史を持っているものが多い。特に、重回帰の理論は正規分布の理論の上に厳密に構築されており、データマイニングのように正規性が担保できないような局面での使用については、問題視する統計学者も多い。しかしながら、回帰分析は要因探索という立場からは極めて利便性が高い。精密標本論をベースに置く数理統計学においては、標本（データ）はデータ生成機序が確率モデルとした母集団からの無作為抽出であることを前提とする。例えば、化学プラントにおける化学製品の収量（y）を目的変数とし、コントロールできる変数、温度（x_1）、総風速度（x_2）、材料の濃度（x_3）、触媒量（x_4）等々を説明変数とする。

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_4 + \varepsilon$$

のような重回帰モデルを作成する。誤差εについては、初等的なモデルでは平均ゼロ、分散σ^2の独立な正規分布を想定する。実際のデータでは、正規分布が成り立つとも限らないが、我々は正規分布が成り立つことを仮定することにより、近似的にでもある種の最適性を満足する要因分析が可能となるというメリットを持っている。

一方、消費者の購買データ（例えば車やコピー機）のように、フィールドで調査した故障記録のデータでは、車の使用時のどんな条件で使用するかの環境条件をコントロールすることは不可能である。もし、補助情報としてどのような環境で使用されていたのかの記録のデータがあれば、それを説明変数として、寿命の予測モデルが作成できるのである。1970年代に提案され、医学・薬学系で多用されている。Coxモデル[5]もそうしたモデルである。

$$\lambda(t;\beta) = \lambda_0(t)\exp(\mathbf{x}^T\beta)$$

ここに、$\lambda(t;\beta)$はハザード関数と呼ばれるものであり、信頼性工学の分野では、瞬間故障率関数とも呼ばれている量である。ここでこのモデルを紹介したいのは、$\lambda_0(t)$というベースラインハザード関数は時間の関数であり、無限次元のパラメーターを持っている。しかしながら、条件付き最尤法の考えに基づき、$\lambda_0(t)$を推定しないでも、個々の要因分析に必要なパラメーターβの推定、検定ができるということである。

ビッグデータの難しさはどこにあるのかというとその難しさは現象を忠実に表現しようとすれば、モデルのパラメーターが観測の時間の経過、あるいは、個体（顧客）IDとともに増加してしまうとうことになる。つまり、どんなデータが多くなろうとも、パラメーターがそれに追従して多くなるという統計の問題は、多くの場合モデリングおよび推定が難しい。

構造パラメーターと撹乱パラメーター

前節で述べた、無限次元のパラメーターを持つ連続関数のベースライン・パラメーターは我々の関心の対象ではなく、要因分析のためのパラメーターβのみが関心の対象であるとするならば、モデルの中のパラメーターは２つに分類されることになる。１つは関心の対象としての構造パラメーター（structural parameter），
他方は、関心はないが、データを生成する上で必要な撹乱パラメーター（nuisance parameter）である。

例えば、学習理論の分野でn人の被験者からなるグループがk種類の能力テストを受験するものとする。第i番目の被験者が第j番目の能力テストについて$y_{ij} = 1$ならば正解、$y_{ij} = 0$ならば不正解とする。このとき、

$$p_j(\theta_i) = \frac{\exp(\theta_i - \alpha_j)}{1 + \exp(\theta_i - \alpha_j)}$$

のように正答確率$p_{ij} = p_j(\theta_i)$にロジスティックモデルを仮定する。このモデルでは、個人の能力を表すパラメーターθ_iと問題の難易度を示すパラメーターα_jの関数として正答確率が表現されていることに注意する。問題の難しさを推定したいのであれば、α_jが構造パラメーター、個人能力パラメーターθ_iが撹乱パラメーターということになる。被験者の数を増やしても構造母数の真値には収束しない（一致推定量ではない）ということが知られている。なるべく真の値に近づける工夫としては、データのサイズに比例させて撹乱パラメーターを増やしてしまうということをしないようにするということである。つまり、なるべく撹乱パラメーターを少なくするということである。１つ

の方法は、撹乱パラメーターの層別である。明確な層化は困難なので、撹乱パラメーターの空間でクラスタリングすることである。実データでのクラスタリングではなく、パラメーター空間でのクラスタリングであるので、その都度モデルパラメーターを推定しながらクラスタリングを行わなければならないので、計算コストは厭ってはならない。

また、ベイズ推定理論も有効であると考えられる。赤池ベイズでは、時間軸上に整列したパラメーターは滑らかに変化するという、いわゆる、「パラメーターの漸進的変化」という仮説を積極的にパラメーター構造の中に取り込むことが可能である。ベイズ推定を利用すると撹乱パラメーターの変動を押さえることができ、その変動の縮小を構造パラメーターの推定における情報の獲得という形で恩恵が受けられるという特徴がある。いずれにせよ、ベイズモデルはより柔軟な制約条件をパラメーター空間に持ち込むことが可能となる。

まとめ

ビッグデータはいま、まさに時代の脚光を浴びている。しかしながら、データさえ集まれば何かビジネスにつながる結果が出てくるのではという誤解さえ生じかねない。何が本質かを見極める目が必要である。どんなにデータがあっても、必要な情報はごくわずかしか含まれていないかも知れない。統計の言葉では、撹乱パラメーターばかり増え、構造パラメーターを推定する情報は極めて少ないという場合もあるということである。これはまさにビッグデータにおける小標本問題という言い方が可能である。

ビッグデータは、莫大な変数について、質的データ、量的データ、テキストデータ、画像データ等あらゆるデータを極めて高いサンプリングレートで時間軸上に並べられた、複数個体（IDの数も極めて多い）のデータである。まずは、データの整理である。統計学の第1歩は、どんなに計算機技術が進歩した今日においても、集計である。統計解析が得意な分野にリスク解析がある。リスクは頻度×損失である。関心のある因子項目の頻度分析が第１歩ということである。データの可視化もこの頻度が容易にわかるような工夫がなされていることが肝要である。

次にモデリングである。ビッグデータにモデリング不要論もあるが、先にも述べたとおり、ビッグデータはビッグであると同時に、スモールでもあるということである。スモールデータには、仮説表現としての統計的モデリングが有効で有り、適切に構築したモデルでは、クラスタリング手法による撹乱パラメーターの層別、比較的制約条件の緩いベイズ推定による、撹乱パラメーターの実質的量の減少化が可能となる。 ◆◆

■参考文献

1）Hand, D. J., et. al.: Principles of Data Mining (Adaptive Computation and Machine Learning), 2001, The MIT Press

2）Huber, P. J.: Massive datasets workshop: four years after, Journal of Computational and Graphical Statistics, 8, 635-652, 1999

3）水田正弘：ビッグデータとは何か、統計学ガイダンス、数学セミナー増刊（日本統計学会＋数学セミナー編集部）、21-25, 2014

4）K. Takeuchi: Hybrid Solid-State Storage System with Storage Class Memory and NAND Flash Memory for Big-Data Application, to be presented in ISCAS,June 2014

5）Cox, D. R.: Regression Models and Life-Table, Biometrika, 34, 2, 187-220, 1972

●筆者紹介

鎌倉稔成：Toshinari Kamakura. 中央大学理工学部教授
産学官連携・知的財産戦略本部長、理工学研究所所長
日本統計学会理事長、統計関連学会理事長

TIPS【基本用語解説】

最尤法
与えられたデータを用いて確率モデルの確率を最大にするようにパラメーターを決める方法。

極限演算
サンプルサイズを無限にしたときの推定量の性質を調べるための演算。

associative array
引数を数字でなく文字列として与えることができるようにした配列。

sparse array
成分がほとんどゼロであるような行列。

ベルヌーイ分布
０を確率１－ｐで、１を確率ｐでとる確率分布。

ベイズ推定理論
パラメーターに先験的事前分布を仮定し、観測データの情報で修正してパラメーターの推測を行う方法。

赤池ベイズ
赤池弘次が1980年代初頭に開発した、ベイズ推定理論と最尤法とを融合し，事前分布の選択方法を与えたABIC規準によるベイズ推論