データをビジネスに活用する実践アナリティクス

<第 2 回>　アナリティクスとデータ解析

梶山 昌之
株式会社ワイハット

1.　ビッグデータは幻滅期？

前回の連載で，ビッグデータはテクノロジのハイプ・サイクルにおける「幻滅期」にあるというお話をしました[1].

ハイプとは誇大広告という意味で，新技術が登場した時は人々の期待が集まるが，期待が高過ぎて，期待に応えることができない状況が次第に明確になり，その後は幻滅期に入るという考え方です.

しかし，幻滅期に入ったときに注意すべきことは，本当にブームが終わったわけではないということです．ビッグデータの場合は幻滅期に入った後も，経営者がデータ解析に関心を持ち始めたという状況は続いています．その解析と活用を支える周辺技術が発達し，今後は安定したデータの分析と活用が実現する状況になることが期待できます.

2.　ビッグデータ関連の用語の整理

ビッグデータに関連してデータ・アナリティクスやデータマイニングなどの用語がよく使われるようになっています．また，「データ解析」や「ビジネスインテリジェンス」という用語は昔から使われていますが，これらの用語との関係もわかりにくくなっています（図 1）.

そこで，ここではビッグデータ関係の用語を整理することにします．　まず，「ビジネスインテリジェンス」という用語は，ビジネス領域で統計学を応用して問題解決をすることを意味します．前回の連載で，ビッグデータの特徴の一つとして非構造化データであることを挙げましたが，データ変換の技術を駆使して構造化データに変換されたデータは，ビジネスインテリジェンスの対象になり得ます.

次に「データマイニング」という用語は，大量のデータから何らかの規則性を発見することを意味します．既存のデータ解析は，例えばコスト最小化などの目標があって，それに対して分析を行うというアプローチを取りますが，データマイニングの場合は何らかの規則性を発見することがアプローチの特徴となります．マイニングは「埋もれた宝を探し出す」ことですが，どのような宝が見つかるかは分からないので，マイニングと呼ぶわけです.

図 1　ビッグデータ関連の用語

「アナリティクス」という言葉の定義は，データの意味を理解して価値を引き出すために数学的ないし統計的に分析することです．米国でアナリティクス（Analytics）の用語は，通常の「データ解析」を意味します．そこで，ビッグデータも対象とするときには，「データ・アナリティクス」という用語が使われています．しかし，日本では「アナリティクス」という用語は「データ解析」とは区別して，ビッグデータも対象とするという意味も含めて使われているようです.

後に「Hadoop/MapReduce」の用語も良く耳にする用語ですが，Hadoop は大量のデータを複数のコンピュータに分散して処理できるオープンソースのプラットフォームです．MapReduce はデータ処理を複数のコンピュータに配置する Map ステップと，Map ステップでの処理結果を集約する Reduce

ステップから成るフレームワークです．ビッグデータの特徴である大規模のデータを高速に処理する必要があるという要求に応える技術と言えます．

3. アナリティクスは従来型のデータも対象

2014 年 5 月，ビッグデータのピーク期から少しかげりが見え始めた頃に，トーマス・H・ダベンポートが『データ・アナリティス 3.0』[2] を出版しました．

3.0 になると，分析の対象となるデータはビッグデータのみではなく，従来型のデータも含むものとしています．また，データ分析は一部の専門家が担当できるだけではだめで，管理者および従業員もデータ分析のスキルを身につけることが必要であると説いています．

また，この本の出版にあたり，監修者まえがきには「流行に乗ってビッグデータの活用そのものを目的とするのではなく，あくまでビジネス上の課題に対するアプローチの一つであるという認識を忘れないことである」と書かれています．

これは，ビッグデータがブームになって，大手の企業を始め誰もがこぞって大容量のデータ処理システムなどを導入して，ビッグデータの分析に投資したが，期待通りの効果が得られていない場合もあったためと考えられます．

Executive Foresight Online という日立のサイトにおいて，2014 年 1 月に，国立情報学研究所の佐藤一郎教授がビッグデータの本質について記述しています[3]．

その中に「スモールデータすらうまく扱えない企業に，ビッグデータを扱えるはずがありません．ビッグデータであれば何でも知見が得られると思うのは幻想であって，まずは手近にある小さいデータを有効活用することから始めたほうが賢明です」という記載があります．

ビッグデータを扱わなければならない状況にある企業はほんの一握りです．しかし，スモールデータは身近にたくさんあると思います．現場が身近なデータを活用するという習慣が根付けばビッグデータも必要に応じて自然に活用できるようになるということだと思います．

統計の基礎知識を身に着けることは，社会生活を営む上で重要という認識から，学校教育でも統計の教育を行うようになりました．

文章から意味を読み取り活用する能力が「リテラシー」ですが，統計リテラシーは統計データから正しく情報を読み取り活用する能力です．データの意味は現場が最もよく知っていますので，現場も統計リテラシーを向上させることが必要です．

4. 衛星の打ち上げコスト予測

さて，ビッグデータではなくスモールデータにも焦点をあてようという話をしましたので，衛星の打ち上げコスト予測のお話をします．衛星の打ち上げコスト予測はどのように行っていると思いますか？

例えば，日本で開発されたあるタイプの衛星で 100kg の衛星の打ち上げコストが 40 億円，150kg の場合が 55 億円という 2 つの実績データが得られているとします．この状況でプロジェクトのマネージャーは，次の打ち上げ計画で 200kg の衛星を打ち上げるコストを予測し，説明しなければなりません．

これに対する答えとして，この 2 点に直線を引いて，そこから 200kg の衛星の打ち上げコストは 70 億円であると予測していいのでしょうか（図 2）．

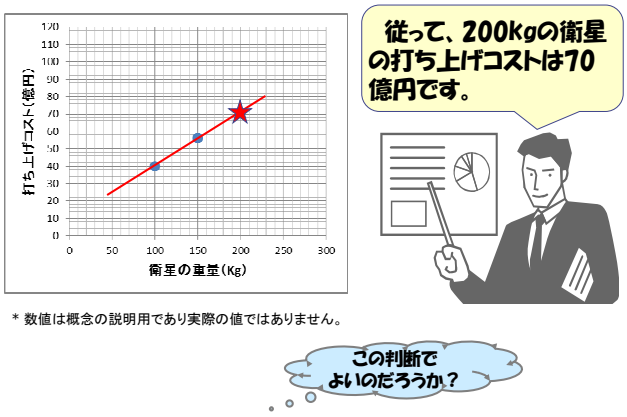

図 2　衛星の打ち上げコスト予測

まず前提となる知識として，2 点では回帰分析の手法は使えません．たとえ 3 点，4 点あったとしても，データが少なすぎて信頼性が非常に低い予測になります．

これを聞くと，半分ぐらいの方は少ないデータで統計的に予測することは無理でしょうと答えます．しかし，対象が衛星なので，何回も打ち上げてデータを取るわけにはいきません．世の中にはデータが取れないものもたくさんあるわけです．その中で予測という行動をしなければなりません．

これが次のテーマであるデータ解析と統計解析の違いになります．先ほどの衛星打ち上げコスト予測の問題の答えは，重量に比例して予測する方式で良いということになります．

理由は，世界のいろいろなタイプの打ち上げ実績から，衛星の打ち上げコストはどのタイプの衛星であっても概ね衛星の重量に比例しているという事実があるためです．その知識を基にこの問題を考えると，たった 2 点であったとしても，今回，日本で打ち上げる 200kg の衛星の打ち上げコストが，2 点を結ぶ直線で予測される値になるだろうと判断するのは合理的です．逆にそういう予測をしなければ，プロジェクトを始められません．

一方，その様な根拠なしに「約 70 億円かかります」と言っても説得力がありません．つまり今持てる情報を最大限に活用して説明することが重要です．

5. 統計解析とデータ解析

次に統計解析とデータ解析の違いについてお話します．図 3 は全体としてはデータ解析の内容を示しています．

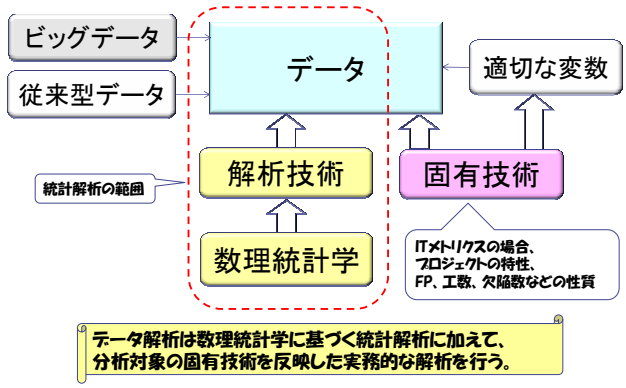

図 3　データ解析（アナリティクス）

まず統計解析ですが，これは数理統計学に基づく解析技術です．例えば正規分布とはどのような分布なのかを明らかにする数理統計学を元に構築された理論により，品質が向上したかどうかを判定するといった話になります．

固有技術とは，例えば IT の場合でしたら，プロジェクトの特性，工数，欠陥数などに関連する知見を意味します．データの分析を行うためにはこれらの性質を理解する必要があります．

固有技術の知識がなければ，適切な変数を選ぶこともできません．例えばソフトウェアの規模を表すメトリクスとして，ファンクションポイント（FP: Function Point）を選ぶのか，ソースコード行数（SLOC: Source Lines Of Code） を選ぶのかを判断する必要があります．計画段階で SLOC を推定することが困難な場合には FP を選択する判断もできるわけです．

データ解析の定義は，数理統計学に基づく統計解析に加えて，分析対象の固有技術を反映した実務的な解析を行うことです．統計解析的に信頼性がないとしても，実務的に合理的な手法であれば，それを活用するというのがデータ解析の立場です．

6. ビッグデータとアナリティクス

ビッグデータはそのままでは分析できないという性質を持っているので，それを分析できる状態にするというプロセスが，ビッグデータを構造化データに変換するプロセスです．

ビッグデータは，パターン認識や大容量データを縮約する方法などの技術が発展すれば，最終的には構造化データになります．そうなれば既存の技術による分析が可能になるということです．

この様に従来型のデータに加え，ビッグデータも解析の対象とするデータ解析をアナリティクスと呼ぶことにします（図 3）．

ただし，ビッグデータを活用することが目的ではなく，問題解決のためにビッグデータの活用が効果的な場合はビッグデータを活用するという考え方になります．従って，従来型のデータを活用したほうが効果的と考えられる場合は，まずは従来型データを分析対象とします．

本連載のタイトルの「実践アナリティクス」は，従来型データに加えてビッグデータも必要に応じて活用する実践的データ解析という意味です．

7. アナリティクスの解析技術

ここでは，アナリティクスの解析技術についてその内容を見ていきましょう（図 4）．

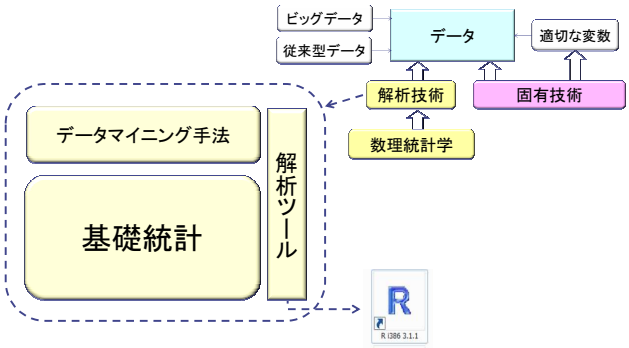

図 4　アナリティクスの解析技術

大きな枠で基礎統計という部分があります．デ

ータマイニング関連の技術を理解するのに，基礎統計の理解が欠かせないことを意味しています．

この部分の習得に一番時間がかかりますが，データサイエンティストとして活動する方には欠かせない部分です．

データマイニングを行う場合，解析ツールとして無料の統計解析プラットフォームである R 言語を使えば，詳しい計算のアルゴリズムを理解できなくても結果は出せます．ただし，結果の意味を正しく解釈し，実務に応用するには基礎統計の知識が必要になってきます．

次にデータマイニングの内容をさらに詳しく見ていくことにしましょう．データマイニングは大量のデータから何らかの規則性を発見することに特徴があると言えます（表 1)．

注意すべき点は，使用される手法は昔から使われているデータ解析の手法とほとんど変わらないということです．

ただし，何らかの規則性の発見というところで，新しく追加されたものもあります．例えばニューラルネットワークは脳の構造を模したアルゴリズムで判断を行う技術ですが，多層構造のニューラルナットワークであるディープラーニングにより近年になって実用化が進んだ技術です．

また，データマイニングは，いつも分析の目標を定めずに分析を行うというわけではなく，設定した仮説を検証するために使われる場合もあります．

データマイニング手法の目的は，さまざまな分類の方法がありますが，ここでは，

① 類似性がある集団に分類する（分類）
② 変数間の関連を調べる（連関）
③ 合否などを判別する（判別）
④ 将来を予測する（予測）

に分類することにします．

分類，連関，判別，予測の列が右側にあります．

表 1　データマイニングの手法と目的

どの手法がどのような目的に役立つかを○印で示しました．

データマイニング手法で，まず習得していただきたいのはクロス集計です．単純ですが非常に強力な手法です．仕事で重回帰分析は使っているがクロス集計は使っていないという方も多い様ですが，クロス集計による層別の検討を行えば分析の方向性が変わるかもしれません．

次回は，プロセス改善とデータ解析の関係について解説の予定ですのでご期待ください．

参考文献

[1] ガートナー: 「ガートナー,「日本におけるテクノロジのハイプ・サイクル：2015 年」を発表」，2015 年 プレス・リリース，2015 年 10 月 27 日， http://www.gartner.co.jp/press/html/pr20151027-01.html.
[2] 佐藤一郎: 「ビッグデータの本質 ビッグデータ活用の鍵を握るのは，現場のインテリジェンス」，日立: Executive Foresight Online, http://www.hitachi.co.jp/products/foresight/strategy/001/, アクセス日: 2016 年 1 月 2 日.
[3] トーマス・H・ダベンポート: 「データ・アナリティクス 3.0」，日経 BP 社, ISBN978-4-822 2-5013-3, 2014.

分析手法	内容	手法の目的				
		分類	連関	判別	予測	その他
クラスター分析	似たものを集める	○				
多次元尺度法	データの構造を調べる	○				
自己組織化マップ	教師なしで学習する	○				
クロス集計	2つの属性の関係を知る	○	○			
主成分分析	多くの変数を少数の変数に要約	○	○			
因子分析	量的データから共通因子を発見する	○	○			
コレスポンデンス分析	カテゴリー間の関係を分析	○	○			
決定木CHID法	質的変数で判別する	○		○	○	
決定木CART法	要因を分析し将来を予測する	○		○	○	
ニューラルネットワーク	入力を判別し分類するモデルを作成する	○		○	○	
アソシエーション分析	頻出するアイテムの組み合わせを発見する		○			
時系列パターン分析	順序性のあるパターンを抽出し予測する		○		○	
線形判別分析	教師データで判別する			○	○	
ロジスティック回帰分析	比率を予測する			○	○	
多項ロジットモデル	3つ以上のカテゴリーの比率を予測する			○	○	
線形回帰分析	目的変数を説明する線形予測式を得る				○	
ダミー変数法	質的変数を含む重回帰分析				○	
ポアソン回帰分析	発生頻度を予測する				○	
階層線形モデル	階層構造をもつデータを分析する				○	
非線形回帰分析	目的変数を説明する非線形予測式を得る				○	
自己相関モデル	時系列データで値を予測する				○	
形態素解析	テキストから名詞と形容詞を抽出する					○
実験計画法	実験を設計し因果関係を明らかにする					○
各種のグラフ	データの可視化と外れ値検出	○	○	○	○	○