大学研究室探訪

滋賀大学

データサイエンス学部

准教授　姫野哲人

1．滋賀大学データサイエンス学部について

滋賀大学データサイエンス学部は 2017 年４月に日本初の学部として開設した。高校生にはデータサイエンスという言葉はまだ聞き馴染みがないかもしれないが，現在，多くの分野においてデータサイエンスは必須のスキルとなっている。

データサイエンスとは，ビッグデータ等の様々なデータを処理，分析し，その結果得られた知見，傾向，特徴をもとに新たな価値を生み出す（様々な課題解決を行う）技術である。これまで勘と経験に基づいて行われてきた判断が，客観的な根拠に基づいて行われることで様々な成果があがっている。例えば，故障予知，良・不良分析，売上予測，要因分析，経路に関する組合せ最適化などの多くのテーマがデータサイエンスの対象となっており，数々の有用な結果が得られている。

データサイエンスのための必要なスキルは大きく分けて，情報学，統計学，ビジネスの３つのスキルであり，本学部ではこれら３つのスキルをそれぞれ身につけることができる。膨大なデータを収集，整理，管理をするためにはデータエンジニアリング力（情報学）のスキルが必要であり，整理したデータを分析するためにデータサイエンス力（情報学および統計学）のスキルが必要となる。これだけで完結しているように思うかもしれないが，実際の課題を解決するためには，そもそもの課題をどのように設定するか，得られた結果から考えられる現実的（実行可能）かつ，意味のある解決策は何かを考える力（ビジネス力）が重要となる（図１）。ビジネス力は一朝一夕で身につくものではなく，様々なデータ分析を重ねること，また，様々な課題解決の事例を知ることが重要となる。滋賀大学では統計学や情報学のスキルを身につけられるだけでなく，ビジネス力を身につけるため，様々なデータ分析を行う実習や，現役データサイエンティストによる講義がある。また，多くのデータサイエンスに関するインターンシップを企業と連携のうえ計画しており，在学中からデータサイエンスの最先端に触れることができる。

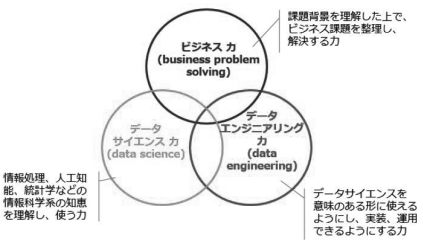

図１　データサイエンティストに求められるスキルセット（一般社団法人データサイエンティスト協会，2014 年 12 月 10 日プレスリリース）

2．研究分野

筆者の主な研究テーマは高次元データ分析である。現在，センサー機器やハードディスク，CPU などの高度化，低価格化が進み，膨大なデータの収集，分析が可能となっている。膨大なデータというのは，単に測定回数が増えているだけではなく，観測項目（変数）も増えている。

一般に変数が増えすぎると，使えなくなる手法があったり，各種手法の精度が落ちたりする。そのため，１つの対応策としては有用な変数を見つけ出し，その変数のみを用いて分析を行うことである。しかし，この方法では少なからず情報を失ってしまうので，価値ある情報が少ない場合には期待した結果が得られなくなることがある。高次元データ分析とは，変数の数が膨大であっても，それらすべてをうまく使って情報を得る方法である。

高次元データ分析の例を一つ紹介する。1000 変数（次元）のデータが２種類（A タイプ，B タイプ）のうちどちらかから得られているとする。A タイプはすべての変数のデータが平均 0.1，分散１の正規分布から得られているとし，B タイプはすべての変数のデータが平均 −0.1，分散１の正規分布から得られているとする。つまり，A タイプのデータは 0.1 を取りやすく，B タイプの

データは -0.1 を取りやすいが，個々の変数だけで見た場合，その差は限りなく小さい（図２）。

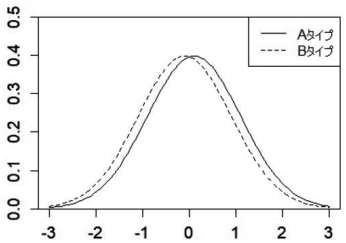

図２　２つのタイプの各変数の分布（密度関数）

このように個々の変数では差が小さい場合でも，それらが数多くあれば精度の高い判別ルールを作成することも可能である。Ａタイプの観測値が 10 個（10 個 ×1000 変数），Ｂタイプの観測値が 10 個（10 個 ×1000 変数）あるとき，それぞれのタイプの各変数の平均の差を計算すると図３のようになる（Ａタイプのほうが平均が大きくなる確率は約 67％である）。

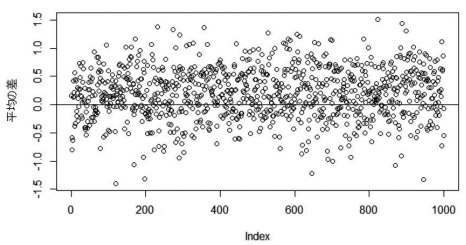

図３　２つのタイプのデータの平均の差

この 1000 次元の平均の差や，各タイプの分散を用いて判別ルールを作成する。新たにＡタイプまたはＢタイプから得られたデータがどちらから得られたかを判断する場合，高次元の手法を用いると高精度で判別可能となる。

ここでは，２つのクラスの判別ルールを作成する問題について紹介したが，高次元データ分析は他にも様々な問題について手法が研究されている。これらの手法は，豊富なデータを扱う現在において活用場面が多い手法である。

３．実データを用いたデータ分析

滋賀大学データサイエンス学部のゼミでは，通常の理系学部のように新たな手法の研究を行うことを主目的とするのではなく，実際の課題に対してデータ分析を行い，課題の解決案を作成する（もちろん，その過程で新たな手法の研究が必要な場合もありえる）。また，本学部の特徴として，大学院（データサイエンス研究科）に所属する大学院生の多くが派遣社会人の現役データサイエンティストであるので，学内で彼らから多くの刺激を受けることも多い。データ分析の経験がない学部生にとって，大学院生からのコメントはとても勉強になるものである。

筆者のゼミではデータ分析の練習用に，人の移動データ（モバイル空間統計），太陽光発電量データおよび気象・大気データ，様々な飲料やお菓子の POS データ，ある製品のセンサーデータおよび良・不良判定データ（連携企業より提供）などが利用できる。これらのデータを用いて分析スキルを鍛えた後，学生にはコンペティションの参加，インターンシップへの参加を推奨している。本原稿執筆時点では，滋賀大学データサイエンス学部の１期生はまだ卒業しておらず，卒業研究の話に触れることはできないが，筆者のゼミの学生（１期生）のこれまでの研究活動について紹介する。

３年生の夏休み時には１人１社以上の中長期（２，３週間）のインターンシップに参加し，データ分析の経験，ビジネス課題の解決案作成の経験を積み重ねた。その後，複数のコンペティションに参加し，これまでに学んできた成果を発揮した。それらのうち，スポーツデータ解析コンペティションでは奨励賞を受賞した。

４年生の前期ではこれまでに培ったデータ分析スキルをもとに，ある製品のセンサーデータおよび良・不良判定データの分析（企業で使用している自動判別の精度を超える判別ルールの作成）を行い，マーケティング分析コンテスト 2020 へ参加している。これらをまとめて卒業研究とする予定である。

４．高大連携について

滋賀大学は多くの高校とデータサイエンスに関する高大連携協定を進めている。たとえば，滋賀県内の高校では彦根東高等学校，虎姫高等学校と高大連携協定を結び，アクティブラーニングに関

する指導・協力を行っている。また，膳所高等学校については野球班のデータ班に対するデータ分析の指導を行った。その分析結果に基づく極端な守備シフトが第 90 回記念選抜高等学校野球大会で効果を生み，多くのメディアから注目を集める結果となった。県外の高校についても香川県立観音寺第一高等学校と連携協力協定を結び，データ分析に関する様々な指導を行っている。

連携協定を結んでいない高校についても，模擬講義の依頼を多く受け，個別に教員が担当している。筆者もいくつかの高校の模擬講義を担当しているが，そのうちの１校である愛媛県立松山南高等学校（以下，松山南高）との連携について紹介する。

松山南高で模擬講義を隔年で行っている。松山南高では，データ分析の課題研究を行っており，データサイエンス関連のコンペティションをいくつか紹介した。このコンペティションに参加するに当たって，生徒が作成したスライドや原稿について様々なアドバイスを行った。主なアドバイスとしては

●適切な分析を行うこと

学んだ手法を列挙するのではなく，目的に応じた適切な手法を選択すること。

●内容を分かりやすくまとめること

テーマが専門的であればあるほど，その分析内容を初めて聞く人に伝えることは難しい。できる限り分かりやすい言葉を使うことが重要となる。

●ストーリーを重視すること

仮説・疑問に応じた可視化・集計を行い，その結果分かったことから新たな疑問を感じれば，新たな分析を実施する。また，自分でまとめた結果については，その妥当性についても評価する。

という３つに関連する部分である（具体的なテーマ決め，分析手法の選択等についてのコメントはほとんど行っていない）。上記で挙げた３つは通常の授業だけでは身につけることが難しく，数多くのデータ分析を行うことで身についていくものである。これらのアドバイスは生徒に向けたものであると同時に，指導する先生方にこれらの視点に注目してほしいという思いもあった。様々なコンペティションのスライド等にアドバイスを続けていくうちにどんどん質が上がっていき，現在では見せてもらう初校原稿にほとんどアドバイスをすることがないほどとなっている。これは，指導する先生方が優れた指導をしているためであろう。これは，これまでに松山南高が数あるコンペティションで次々に入賞していることが物語っている（地域創生政策アイデアコンテスト 2018，2018 年度・2019 年度統計データ分析コンペティション，第８回・第９回スポーツデータ解析コンペティション，第９回データビジネス創造コンテスト，第３回和歌山県データ利活用コンペティション等多数）。

筆者は高校生のデータ分析にアドバイスを行う際，大学で学ぶ高度な分析手法を紹介することは避けている。高度な分析手法を使えば，データから得られる情報は飛躍的に増え，生徒のスキルも上がり，データサイエンスに関する興味も深まるだろう。また，うまく使いこなすことでコンペティション入賞の確率も上がるだろう。ただし，（以下は個人的な意見となるが，）高度な分析手法を使ってコンペティションに入賞することは，入賞した生徒にとっては有益かもしれないが，その他大勢の高校生にとって悪影響を与える可能性がある。それは，コンペティション参加を考えていた生徒に対し，高度な手法を用いなければならないという印象を与えてしまうと，参加をあきらめ，データ分析の経験を失ってしまうかもしれない。また，多くの高校生が現在学んでいる基本的なデータ分析手法をつまらないものと感じてしまい，データサイエンスに対する興味を失ってしまうかもしれない。

高校で学ぶデータ集計，可視化の知識だけでも，データから十分な知見を得ることができるということを高校生にも高校の先生方にも知ってもらいたい。そのためにも，数あるコンペティションに目を向け，興味のあるコンペティションがあれば是非とも積極的に挑戦してもらえればと思う。