

# ビッグデータからの仮説導出のための視覚化

## Visualization for Data based Hypothesis Building

共同研究メンバー

○今泉忠\*、佐藤洋行\*、久保田貴文\*（○代表、執筆者）

**Keywords** : Big Data, Clustering, Interaction Terms

### 1. 研究目的

実際のビジネス場面でビッグデータを利活用して、ビジネス課題解決に役立てることが日常的に行われている。例えば、佐藤は多摩丘陵地に造成され近年高齢化が進んでいる多摩ニュータウン地域に居住する消費者について、その消費行動パターンから多摩ニュータウンの街造りの現在と未来について検討するために、消費者の購買行動に関する ID-POS データをもとにした分析を行っている。また、久保田はビッグデータについて空間的な特性を踏まえた統計的モデルについて研究している。いずれにしても機械学習で発達してきた手法を適切に適用して、課題解決のために結果を活用する必要がある。佐藤による研究の場合では、対象集団が主として多摩市に居住する住民を想定しており、その情報をもとにして特性を反映した仮説検証型アプローチを用いた分析が可能であり、実践的な課題解決に役立てることができると考えられる。しかし、データに関する事前情報が少ない場合には、データから探索的に因果に関する仮説導出型アプローチが有用であることがある。この場合に、分析者が納得できるように提案された仮説（モデル）について理解を支援することが望ましい。本研究ではこの理解支援に関しての分析ステップを提案する。

### 2. 研究概要

#### 2.1 仮説について検討するための目的変数に関する分布

本研究では、教師ありの場合に、データにもとづいて仮説発見型アプローチを行う場合を扱う。まず、教師である目的変数については、基礎的な正規分布  $Y \sim N(\mu, \sigma^2)$ 、二項分布  $Y \sim B(n, p)$ 、ポアソン分布  $Y \sim Po(\lambda)$ 、多項分布  $P(n_1, \dots, n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}$  に限る。

#### 2.2 説明変数によるグルーピング

データが複数のグループからなっており、グループについての識別情報がある場合の判別や分類の問題を扱う主な手法としては、線形判別、2次判別、ロジスティック判別、多項ロジス

\* 多摩大学経営情報学部 School of Management and Information Sciences, Tama University

ディック判別、決定木・分類木、ナイーブベイズ、サポートベクトルマシン、ランダムフォレスト、k-means の結果にもとづく分類などがあり、これらの手法の比較研究なども行われている。しかし、本研究の目的は、仮説発見型アプローチであるので、このようなグループ分類を行うこと自体が仮説発見となる場合もある。したがって、分析結果をもとに、分析者に仮説発見（モデル発見）を促すために複数のモデルを比較し、何らかの形でその根拠を示せるようにする。このような場合に、モデル選択の指標としては情報量基準関連の AIC や BIC などが提案され、また決定係数  $R^2$  や F 比などもよく知られている。ここでは、結果から仮説発見を試みることを考えているので、ある統計モデルを想定した場合に、予測値を用いて、視覚的に表示して仮説について検討できるようにする。

いま、あるデータ  $(y_i, x_{i1}, \dots, x_{ip})$  をもとにモデル  $M_1$  とモデル  $M_2$  を設定し、比  $K_i$  を

$$K_i = p(y_i|M_2)/p(y_i|M_1)$$

として定義して、この値が 1 より大きい小さいかで、どちらのモデルが確からしいかを示す指標として用いることにする。本研究ではベイズ流のモデル構成を行っていないので、これはベイズファクターではなく、モデルの確かさを検討することはできないが、どのような仮説が想定できるかについて示唆を与えるために、各データ  $i$  について比  $K_i$  を求め視覚的に表示する。

### 2.3 説明変数の組

2.2 で述べたように  $N$  個のケースに関するデータは  $G$  個のグループに分類することができることも考えられる。この場合に探索的に仮説を探するためのモデルとしては、少なくとも以下の 2 つが考えられる。

モデル  $M_1$ ：説明変数の組として  $(x_1, x_2, \dots, x_p)$  の主効果モデル

モデル  $M_2$ ：説明変数の組としてグループ  $g$  毎の  $(x_{g1}, x_{g2}, \dots, x_{gp})$  とした交互作用モデル  
したがって、分析者は次のステップを行うことで仮説に関する比較を試行錯誤的に行うことができる

Step 1：説明変数の組を設定する。

Step 2：必要ならば、変数について標準化などの変数変換を行い、変換後の変数の組を説明変数とする。

Step 3：k-means などのクラスタリング手法を適用して、 $G$  個のグループを作成する。

Step 4：目的変数に関する分布を設定する

Step 5：目的変数に対して、説明変数およびクラスター所属変数を用いて 2 つのモデルモデル  $M_1$  とモデル  $M_2$  を当てはめる。

Step 6：比  $K_i$  を求めて視覚的に表示する。

## 3. 応用

株式会社マーケティング・サービス社からご提供いただいた車の購入に関するデータに適用した。このデータはマーケティング調査専門会社による調査であり、調査項目の選定などが適切と考えたことによる。インターネット調査によりデータ収集を行った。データの大きさは 462 であった。マーケティングのデータの分析で顧客満足度を目的変数とする場合もあるが、実際の購入と満足度の関係がどのように対応しているか明確ではない点がある。そこで本研究

では、1 台前の車メーカーから現在所有している車メーカーへの推移を個人毎のロイヤルティと解釈し、そのデータから仮説を発見することにする。

Step 1: 車メーカーとしては、トヨタ、日産、...、BMW、その他の外国車からなる 12 メーカーとした。目的変数は、現在所有している車のメーカー（質的変数）であり、説明変数としては 1 台前に所有していた車メーカー（質的変数）と、5 点評定尺度で記録した車所有に関する 10 項目

- |                         |                     |
|-------------------------|---------------------|
| a. 洗車には十分時間を掛けており洗車が好き  | b. 車に乗ると他人の目が気になる   |
| c. 他人と違った車に乗りたい         | d. 最新のモデルに乗りたい      |
| e. 車は仕事や生活に必要なだから乗っている  | f. 再販価格を考えて車を選ぶ     |
| g. 街でよく見かけるモデル（車）を選ぶ    | h. 扱いやすい、運転しやすい車を選ぶ |
| i. 機能やデザインより本体価格の安い車で十分 | j. 車を運転するのがとても好き    |

である。

Step 2: 1 台前に所有していた車メーカーについて 2 値のダミー変数に変換し、また、項目 a~j については 4 点以上を 1、それ以下を 0 となるように変換した。

Step 3: 変換した変数 a~j についてクラスター分析（k-means）を適用した。最終的には G = 4 のクラスター解を採用した。

Step 4: 現在所有している車のメーカーについて多項分布を仮定した。

Step 5: 説明変数として、2 値のダミー変数と項目 a~j の変数を用いて、3 つのモデルを当てはめた。

モデル  $M_1$ : 説明変数の組として 2 値のダミー変数のみを用いたモデル

モデル  $M_2$ : 説明変数の組として 2 値のダミー変数と項目 a~j の変数のみを用いたモデル

モデル  $M_3$ : 説明変数の組として 2 値のダミー変数とグループごとに項目 a~j の変数への係数が異なるとしたモデル

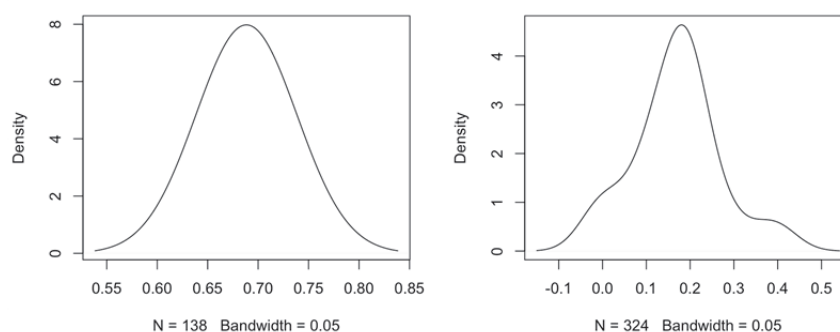


図 1: 現在の車メーカーがトヨタ車に推移した場合に、モデル  $M_1$  で 1 台前の車メーカーがトヨタであった場合（左側）とそうでない場合（右側）に推定された密度

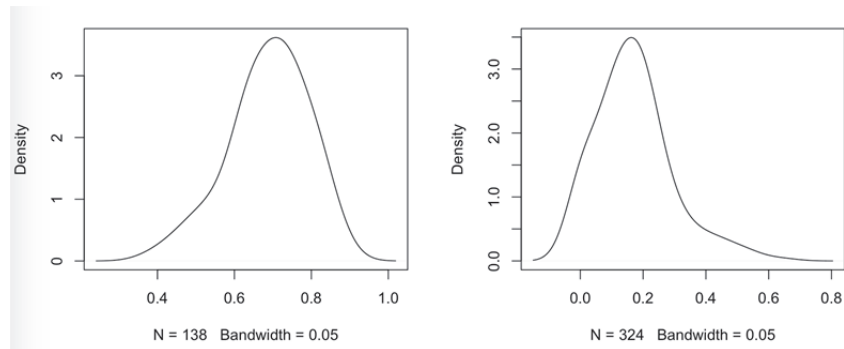


図 2：現在の車メーカーがトヨタ車に推移した場合に、モデル M<sub>2</sub> で 1 台前の車メーカーがトヨタであった場合（左側）とそうでない場合（右側）に推定された密度

車メーカーがトヨタである場合には、前回の車メーカーがトヨタであるかどうかは現在購入した車メーカーを決定しているのでないかとの仮説設定が示唆される。

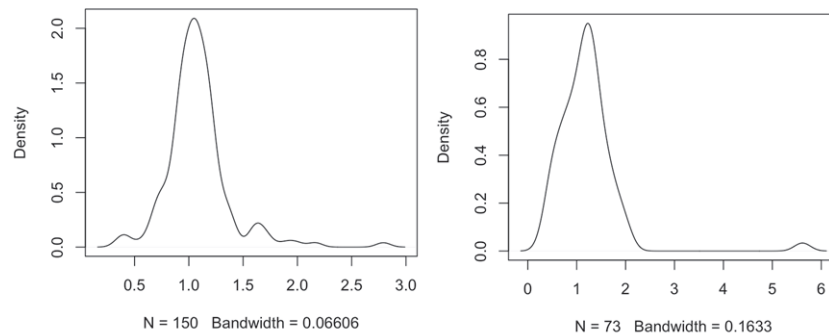


図 3 現在の車メーカーがトヨタである場合（左側）とホンダ車（右側）での比  $K_i$  の分布

図 3 からはトヨタ車はメーカー優位でホンダ車はメーカーおよび車への選好が車購入に影響しているのではないかと示唆される。

#### 4. 総括

データから仮説導出を行う場合について、モデルについて、個別のデータで比較できる方法について検討した。しかし、データが無作為に抽出されている場合でも、対象集団やそのバックグラウンドがどのように反映しているかどうか不明であるので、分析者が関与していないところで、扱っているデータが偏っている可能性がある。それを踏まえた方法の検討が必要である。