研究レポート　No.14　～bodais スコアリングのカテゴリ統合～

2020 年 12 月 2 日　株式会社アイズファクトリー　https://bodais.com/company/

概要

カテゴリカルな説明変数の水準数が過剰になっているデータに対して、適切に水準数を削減する技術をカテゴリ統合と呼ぶ。過剰な水準数に起因して、カテゴリごとの説明能力が低下している場合であっても、カテゴリ統合により安定したモデル構築が実現する可能性がある。本稿では、bodais のスコアリングエンジンに搭載されているカテゴリ統合機能に関連して、カテゴリ正解率とカテゴリ統合の判定指標について解説する。

1.　はじめに

bodais のスコアリング分析では、正解として 0 又は 1 の値を目的変数としたロジスティク回帰 [1] の学習モデルを構築する。ロジスティク回帰では通常、カテゴリカルな説明変数を解析対象とするが、 bodais ではデータ変換アプリ[2]を通じて、連続値にも対応できる。一般的に、データ解析を行う際には、解析の精度を高めるため「変数選択」「変数加工」などのクレンジング処理を手動で行う必要があるが、bodais には様々なクレンジング処理が実装されており、カテゴリ統合処理、欠損処理、マルチコ処理等を行う（オートクレンジング機能）。また、予測時には、モデル作成時の説明変数に存在しないカテゴリ値がある場合でも、適切な前処理により計算エラーを起こさずに予測値を算出する処理が組み込まれている。これら様々な前処理を経てスコアリング分析が行われる。

カテゴリ統合処理では、ある一つの説明変数におけるカテゴリの種類の数（水準数）が多すぎる場合には、水準数を調整することによって分析精度の向上を図る。この調整は学習データにおける正例（値 1 の正解）の割合（以下、「正解率」と呼ぶ）をコントロールすることで行われる。本稿では、正解率とサンプル数の空間における、データの相構造について解説する。

2.　カテゴリ正解率

図 1 に bodais スコアリング学習時に使用するデータ例を示す。1 列目は顧客会員番号（ID）、5 列目は目的変数（正解）である。2~4 列目は説明変数（性別、年齢、住所）である。

一般的に使用される正解率$\bar{p}$は、式(1)で定義される。この正解率$\bar{p}$を全体正解率と呼ぶことにする。

$$\bar{p} = \frac{M}{N} \qquad (1)$$

ここで、Mは正例数（正解列が 1 である行数）、Nはサンプル数（総行数）である。全体正解率$\bar{p}$は、説明変数のカテゴリ値に依存しない形だが、この概念を説明変数のカテゴリごとに定めることができ、次の様にカテゴリ正解率 $\bar{p}(C)$ を定義する。

$$\bar{p}(C) = \frac{M(C)}{N(C)} \qquad (2)$$

ここで、$N(C)$はある説明変数のカテゴリがCである場合のサンプル数、$M(C)$はカテゴリがCである場合の正解数である。具体的には、図 1 における年齢のカテゴリ 2 の場合、$N(C) = 2,\ M(C) = 2,\ \bar{p}(C) = \frac{2}{2}$ =1.00 である。同様に、性別のカテゴリ 0 の場合、$N(C) = 4,\ M(C) = 3,\ \bar{p}(C) = \frac{3}{4}$ =0.75 である。ここで、カテゴリ正解率$\bar{p}(C)$が 0.5 以上の場合には、$\bar{p}(C)$→1- $\bar{p}(C)$の変換を通じて、議論の対称性が成立するため、カテゴリ正解率$\bar{p}(C)$が 0.5 未満の領域のみ考察すれば十分である。そのため、以下、$\bar{p}(C)$<0.5 とする。

ID	性別	年齢	住所	正解
1	0	2	16	1
2	1	1	12	0
3	0	4	30	1
4	0	3	9	0
5	0	2	16	1

図 1. bodais スコアリングの入力データの例。赤枠は年齢のカテゴリ値が 2 のみ抽出した場合を示す。

3.　カテゴリ統合

式(2)の定義から明らかなように、カテゴリ正解率 $\bar{p}(C)$ と全体正解率$\bar{p}$が大きく異なることが発生しうる。特に、$N(C)$と$M(C)$のサンプル集合の相関が弱い場合や、ある説明変数の水準数が多い場合には、カテゴリCの説明能力が低いことがある。ここで、カテゴリの説明能力はカテゴリ正解率で測ることができる。例えば、図 1 における年齢のカテゴリ 2 は 2 個存在し、ともに正解が 1 である（$\bar{p}(C) = 1.0$）。よって、カテゴリが 2 であれば正解 1 であることを 100% 説明できる。仮に、カテゴリ 2 の正解の一つが 1 でなく 0 である場合、カテゴリが 2 であっても正解が 1 と 0 の両者が存在するため説明力は前者よりも低下する（$\bar{p}(C) = 0.5$）。このように正解説明能力の低いカテゴリを特定し、他のカテゴリと結合することで、全体として説明力の調整を行う。

通常のデータ分析では、多すぎる水準数の分類粒度をデータの特性により調整する。例えば、年齢であれば、年代で分けるなどの処理を行う。bodais のカテゴリ統合はこの人手で行っている処理を自動で行っているイメージに近いことを補足しておく。

4.　カテゴリ統合の判定領域

カテゴリ統合時には、カテゴリ統合をするかしないかの判定が重要になる。bodais では単純にカテゴリ正解率のみで判定せず、下記に記すような複雑な処理により判定を行っている。

統計精度を担保するのに必要なサンプル数nは、信頼幅 L 、母集団の分散 σ^2とする式(3)で記載されることがよく知られている [3]。

$$\mathrm{n} = \frac{N}{\frac{N-1}{\sigma^2}\left(\frac{L}{2k_\alpha}\right)+1} \qquad (3)$$

bodais では、式(3)を応用して、カテゴリ正解率 $\bar{p}(C)$に対するカテゴリ統合の判定指標 β を導いている。 β の値が大きい極限は、全てのカテゴリを統合対象としてしまう過統合を

示唆し、小さな値は統合の必要性が希薄化する傾向を表す。モデルの挙動はβの値とカテゴリ正解率、サンプル数の値の領域に応じて様々な傾向を示すため、各領域の特徴に応じて、カテゴリ統合の可否判断と統合処理の詳細を決めていく。図2は、横軸にカテゴリ正解率$\bar{p}(C)$、縦軸にサンプル数Nの二次元面におけるβの分布をカラースケールによって示したものである。

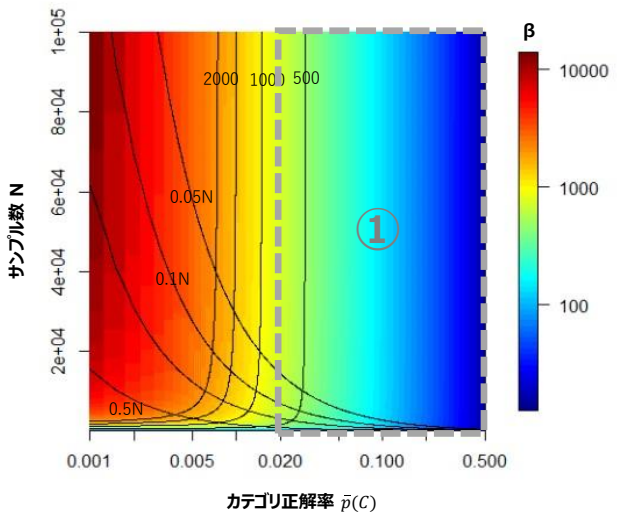

図2. カテゴリ統合判定指標の分布図。横軸はカテゴリ正解率$\bar{p}(C)$、縦軸はサンプル数N。判定指標βを色で表示。

図2の左半分にあたるカテゴリ正解率$\bar{p}(C)$が低い領域では、判定指標βが大きな値（赤～黄色）をとっており、過統合の可能性を抱えているため、注意が必要な領域である。図2の右半分のβが小さい領域（$\bar{p}(C)$が2%以上の①の領域）では、カテゴリ正解率$\bar{p}(C)$が高いため、サンプル数Nが十分あればカテゴリ統合などの処理は必要性が薄れる。しかし、Nが十分でないと過学習の可能性があるためカテゴリを整理する必要がある。

カテゴリ正解率$\bar{p}(C)$が2%未満の領域については、図3（対数グラフ）に見るように、サンプル数Nの大小による違いが存在するので、2つの領域（②、③）に分割して考える。

図3は図2の判定指標βの代わりにサンプル数Nで除算した値(β/N)で表示した。図3の②の領域($\bar{p}(C)$が2%未満、曲線（0.1N）より下の領域）では、βが非常に大きい。Nも$\bar{p}(C)$も共に小さい左下の隅を中心に問題が起きやすい危険な領域である。この領域では、予測誤差が大きく、学習が十分にできない可能性がある。存在意義の小さい細小なカテゴリを削減して変数を簡略化するほうが良く、サンプル数に応じた閾値の調整を行うといった例外処理を入れる工夫が必要である。

一方、③の領域（$\bar{p}(C)$が2%未満で、曲線（0.1N）より上の領域）では、βが大きくなく、Nが適度に大きいため、カテゴリ統合は適度に働く。

図2,3上に、カテゴリごとに該当する点を表示させることができる。例えば、図3中の白丸（●）は、あるデータにおいて、カテゴリ統合が発生したカテゴリのうち2つの例を示す。このデータでbodaisスコアリング解析をしたときのゲインチャート[4]を図4に示す。モデル評価値が0.86とよい分析結果であることがわる。

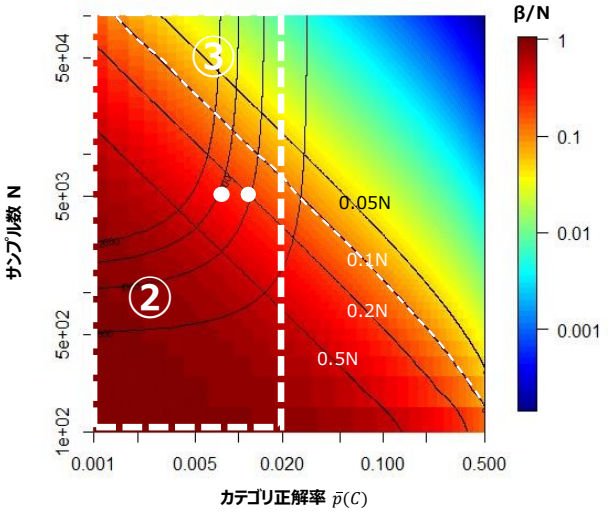

図3. カテゴリ統合判定指標の分布図。横軸はカテゴリ正解率$\bar{p}(C)$、縦軸はサンプル数N。β/Nを色で表示。

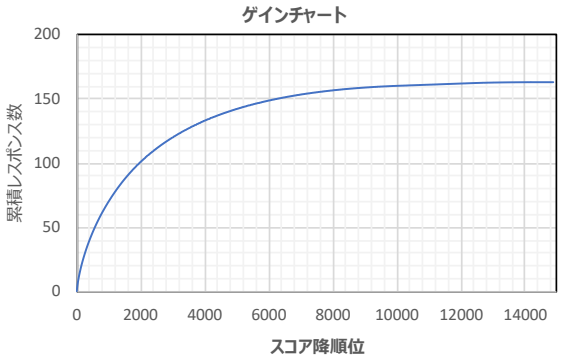

図4. 全体正解率$\bar{p}$=1.09%、サンプル数N=14886の事例。スコアリングのモデル評価値は0.86。このデータでカテゴリ統合が発生したカテゴリの一部を図3の白丸（●）で表示。

5. まとめ

本稿ではbodaisのスコアリングの前処理として行っているカテゴリ統合処理に関連して、判定指標の分布について領域ごとの特徴を記述した。この特徴の理解は、カテゴリ統合を行う理由やカテゴリ統合の発生・非発生の定性的分析に役立つことが期待される。

6. 参考文献

[1] ロジスティク回帰分析, 株式会社アイズファクトリー, 研究レポート No.3 (2011).
[2] https://bodais.com/system/product/data-converting-app/
[3] 船津好明, 調査統計入門 －単純任意抽出法を中心として－, 共立出版, 166 (1977).
[4] ROC 解析によるモデル精度評価, 株式会社アイズファクトリー, 研究レポート No.2 (2011).