リスク要因の探索と剪定について

リスク解析戦略研究センター 金融・保険リスク研究グループ
准教授　　川崎　能典

1　はじめに

与信スコアリングに代表される信用リスク分析では，判別分析や離散選択モデルは基本的なツールである．昨今はサポートベクターマシンに代表される非線形判別を扱う手法の発展が著しい．カーネルトリックの妙味は，観測値の空間で細かな非線形モデリングをする必要がないところにあるが，実際にリスクモデルを運用する現場では，何が決定要因であるのかを知りたい，変数の顔を見たいという欲求は，専門分野での知見の深化という観点から依然根強いと思われる．

本研究では，カーネル法からすると愚直にも見える「リスク因子の同定」という経路を取って，ロジット・プロビット等の説明変数として有効なクロス変数を探索的に求める，計算統計的なアルゴリズムを提案する．さしあたり本報告では，説明変数の 2 次交互作用項を総当たり的に探すことを目的とするが，原理的にはより高次も探索可能である．

2　分析手順

ここで，関心ある被説明変数は 2 値であるとする．すなわち，陽性・陰性，事故・非事故，購入・非購入，倒産・存続等の 2 値カテゴリー変数である．当面具体的な応用例が登場するまでは，仮に陽性 (= 1)・陰性 (= 0) という言葉で説明しておこう．最終的には主効果 + 交互作用の形に持ち込むとしても，まずは主効果の分析から始めなければならない．ここでは坂元 (1983) に依拠し，多項分布のあてはめモデルを適用することからスタートし，以下の手順に従う．

1. 2 次元クロス表解析
2. 3 次元クロス表解析のためのデータセット作成
3. 3 次元クロス表解析
4. クロス項目の剪定による有望な変数の絞り込み
5. 上記変数に対応するダミー観測値の作成
6. ロジットモデルの推定と確率予測
7. 模擬判別 (アウトサンプル予測)：可能なら

1., 3. では CATDAP(Katsura and Sakamoto, 1983) を利用する．1. は，もし手元のデータが膨大な調査項目を含むような場合には，1 変量として優れている変数の組合せのみで以降のステップに進む，という意思決定をする場合は重要である．データセットが小規模であれば不要とも言えるが，1 変量の特性を眺めておくことは常に有益である．

2. は 3. の準備である．ここではデータセットを組合せ的に生成した上で逐次解析し，結果を収集する．データセットを作成すると同時に，CATDAP に読み込ませるためのデータ記述ファイル

を (小分けにしたデータセットごとに) 作成しなければならない．ここでは，データの生成も，データ記述ファイルの生成も，プログラムで自動的に行っている．

3. は 2. で生成したデータを CATDAP で解析し，必要な出力をファイルで残しておくようにする．ひとつのデータセット (組合せ) につき比較されるモデルは 4 つしかない．3 次元クロス表か，説明変数のどちらかひとつだけを採用した 2 つの 2 次元クロス表モデルか，あるいは独立モデルかである．当然，3 次元クロスが一番良いと判断された組のみが考察の対象になる．

この段階で，連続変数を元にしたカテゴリー変数の生成や，カテゴリーの併合等 CATDAP の中で実現されている機能が大いに役立つことを強調しておきたい．また，多数の名義尺度から構成される調査項目があった場合，予備解析の結果をもとにしてリスクの順で並べ替えを行うなどの操作も，解釈しやすい結果を得るためにしばしば重要である．

クロス表が含むカテゴリーの組合せの中で，判別に効きそうな変数というのは，それらのカテゴリー値を固定した時の目的変数の分布が，データ全体での目的変数の分布と大きく異なる場合であると思われる．例えば，データ全体で陽性が 40% ，陰性が 60% のとき，ある変数 I_0 の値が i_0 で，もうひとつの変数 I_1 の値が i_1 のとき，陽性が 90% で陰性が 10% であれば，このクロス項目は，陽性の判定に一役買いそうな気がする．ただし，このクロス項目に対応する周辺度数が標本全体の数に比してあまりに少なければ，全体的な判別力の強化には，あまり役立たないだろう．

以上の考察に鑑みると，クロス項目の中から,「データ全体での目的変数の周辺分布と陽性・陰性の出現比率が異なる」とデータから判定された項目だけを拾って候補とし，他は剪定してしまうのがよいだろう．これは母比率の検定 (二項検定) によって簡単に行える．プログラム上は，3. が終わった段階で自然と 4. も実現するよう設計している．

5. はやや煩雑なプログラミングを要する．どの変数どうしの，どのカテゴリーの組合せが残ったか，連続型変数の場合は，階級区間を定める数値の情報などを決まった形式で CATDAP から出力させるようにコードを書き換え，出力ファイルにアクセスしてダミー変数を半自動的に生成するようなプログラムを用意している．以上，1. から 5. までは，実際にはスクリプトによって一括して処理される．

あとは候補変数をロジットモデルの右辺に投入するだけである．すなわち，$Y_1 = 1$, $Y_2 = 0$, $Y_3 = 0, \ldots$ というようにデータが与えられていて，各 i に対し X_i という説明変数を用意したとして，F を平均ゼロのロジスティック分布の累積密度関数としたときの，次の対数尤度

$$\ln L(\alpha, \beta) = \sum_i [Y_i \ln\{1 - F(-\alpha - \beta' X_i)\} + (1 - Y_i) \ln F(-\alpha - \beta' X_i)]$$

を最大化する α, β を求める．データが利用可能なら，アウトサンプル予測 (7.) を行う．

このようなリスク要因の探索法はさまざま分野で適用可能であるが，これまでのところ，医学データ，信用リスクの分析等に応用を試みた．大規模な業務データになるほど高次交互作用の相関は高く，上に述べた二項検定による剪定でもなお多くの候補が残るにもかかわらずその殆どが有意にならない．このような状況では，L^1 正則化との併用を考えるのがよいであろう．

参考文献

Katsura, K. and Sakamoto, Y. (1983). CATDAP: A categorical data analysis program package, *Computer Science Monographs No. 14*, The Institute of Statistical Mathematics, Tokyo.

Kawasaki, Y. (2009). Searching and pruning risk factors in the logit mode, *Society for Risk Analysis 2009 Annual Meeting*, Renaissance Baltimore Harborplace, Baltimore, U.S.A.

坂元慶行 (1985). カテゴリカルデータのモデル分析, 共立出版.