非線形分類法を用いた高次元データの変数選択
～健康診断データでの実践～

只野英恵　(指導教員：吉田裕亮)

1　はじめに

日頃, 私達は価値ある判断を得るため, 数ある情報の中から有効なデータを抽出し, 処理を行っている. その例がマーケティングや医療分野での分析, 判断である. 特に多変量のデータから効率的に決定的要因を知ることは重要となる. そこで, 本研究では非線形分類法を用いて, 複数の変量から判別に有効な変量を推定する手法を提案する.

本研究で扱うデータは, 2 群判別がなされた多変量データで, 判定に用いられた変量に, 関連がありそうな変量を加えたものである. この多変量データから非線形分類法を用いて変数削減を行うと同時に誤判別率を求めることにより, 手法の有効性をみる. また, 最終的に要因となる変量の判定に用いられた基準値を求めるところまで検討する.

以下に, 本研究で用いる基本的なツールの概説をまず述べておく.

2　AIC を用いた変数選択

2.1　AIC(Akaike's Information Criterion)

一般的に, あるデータを統計的に説明するモデルを作成することを考えると, 説明変数や次数を増やせば増やすほど, データとの適合度を高めることができる. しかし, その反面, ノイズなどの偶発的変動も取り込んで推定してしまうため, 過適合問題が起こる. そこで, 適合度を見つつ説明変数の数を減らしていくことが重要であり, この適合度を統計量として表した指標として, AIC（Akaike's Information Criterion）が知られている.

$L(\hat{\theta})$ をモデルの最大対数尤度, k を自由パラメータの数とするとき, AIC は

$$AIC = -2L(\hat{\theta}) + 2k$$

で与えられる. 各モデルごとに算出される AIC の値が最小となるモデルが, より最適なモデルであることを意味する.

2.2　変数選択

まず AIC を用いて, 変数選択を行っていく. 変数選択には, 変数増加法, 変数減少法, 変数増減法などいくつかの方法が存在する. 中でもステップワイズ法とも呼ばれる変数増減法は, クラスラベルの回帰式を設定し, そこに AIC 値を最も改善させる変数をひとつ追加し, 一度削った変数も採用することを許すことにより柔軟な結果を返す手法である.

3　カーネル PCA

3.1　PCA(主成分分析)

PCA(主成分分析) とは, 高次元データから互いに独立な成分を推定し, 観測データをそれらの成分の線形結合で説明するものである.

PCA(線形) の欠点は, 非線形なデータの構造が捉えにくいということである. 実際, 複雑なデータに対しては, 線形な構造だけを見ていては不十分なことも多い.

3.2　カーネル法

カーネル法とは, カーネル関数を使用し, 観測データを高次元 (一般には無限次元) のベクトル空間に写像し, 変換後のデータに線形の手法を用いることで, 非線形の関係を考慮することができる. カーネル関数によって, 写像された空間は, 再生核ヒルベルト空間の性質を持ち, 計算の複雑度を抑えつつ, 内積に基づく線形解析手法を高次元ベクトル空間へ拡張し, 実質的に非線形な解析を行うことができる. このことは, 一般的にカーネルトリックと呼ばれる. すなわち, カーネル関数 K を用いて.

$$\langle \Phi(x), \Phi(y) \rangle = K(x, y)$$

と考えることで, 非線形写像 Φ を与えるのではなく, カーネル関数 K を与えることにより, 非線形な手法に変換することが可能となる. よく用いられるカーネル関数として, 線形カーネル, ガウスカーネル、多項式カーネル、ラプラシアンカーネルが挙げられる. 本研究では, ガウスカーネルを用いることとする.

ガウスカーネル

$$K(x, y) = \exp\left(-\|x - y\|^2/\sigma\right), \ (\sigma > 0).$$

3.3　カーネル PCA

カーネル PCA とは, 3.2 で示したようなカーネル関数を利用することで, 線形手法である PCA を非線形化し, データの非線形な方向での分散の大きな成分を抽出することができる手法である. カーネル PCA においても, 線形 PCA のときと同様に主成分を求めることができるので, 本研究では第 3 主成分まで求め, 主成分プロットを行い, 適していると考えられるカーネル関数およびパラメータの値を見つけることにする.

4　実データへの応用

4.1　健康診断データ

約 60 万人の成人女性の健康診断データから 1 万人分のランダムサンプルされたデータを用い, 脂質代謝の判定が A から C, D から F であるものをそれぞれ約 5 百人ずつ抽出し繋げたデータで実践する. このとき, 健康診断項目の年齢, 身長, 体重, BMI, 最高血圧, 最低血圧, 総コレステロール (T-Cho), 善玉コレステロール (HDL), 中性脂肪 (TG), 赤血球数, 白血球数, ヘマトクリット (Ht), ヘモグロビン (Hb), 尿蛋白, 尿糖値, 血糖値の 16 変量をもつ多変量データとする.

4.2　非線形分類法の適用

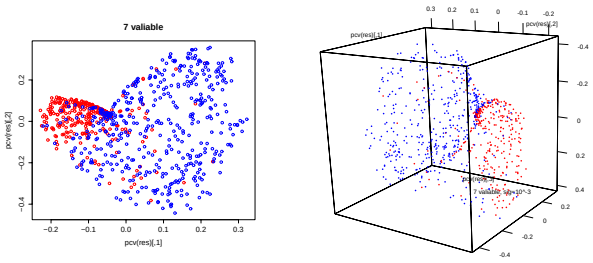
AIC による変数選択を実データに施すと, 2 値判定により, 元の 16 変量から半分以下の 7 変量,「年齢, 体

重, T-Cho, HDL, TG, 尿蛋白, 血糖値」を選択することができた.

ここで, 別の手法を施すことにより変数を更に減らすことができないか検討する. AIC では変数を選択することで次元を下げるという目的を果たしたが, 次に先に述べた, 変数を合成し情報を圧縮するカーネル PCA という手法を適用する. カーネル PCA では, 2 次元ないし 3 次元の可視化が可能となるため, ある変量を除いても 2 群判別への分かれ方にさほど変化が見られない場合, その変数は判別に大きな影響を与えないと考えられる.

まず, AIC で選択された 7 変量から始め, 適宜, 変数を減らし, 比較を行う.

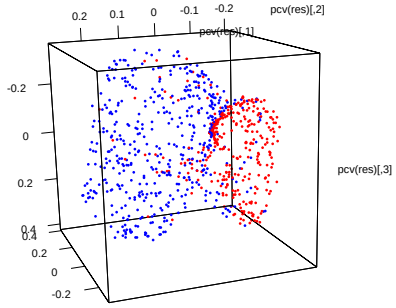
図 1: 7 成分の結果



(2 次元)　　(3 次元)

このとき, AIC の統計的有用性の指標に基づき, 変数を減らしていく. 一例として, 血糖値, 尿蛋白成分を取り除いた 3 次元 PCA プロットが下図である.

図 2: 5 成分 [HDL・年齢・体重・TG・T-Cho]



4.3　誤判別率

表 1: 誤判別率

変数	変数名	誤判別率
7	T-Cho, TG, 体重, 年齢, HDL, 尿蛋白, 血糖値	7.23 %
6	T-Cho, TG, 体重, 年齢, HDL, 尿蛋白	8.16 %
5	T-Cho, TG, 体重, 年齢, HDL	8.16 %
4	T-Cho, TG, 体重, 年齢	9.09 %
3	T-Cho, TG, 体重	9.32 %
2	T-Cho, TG	9.32 %

A から C 判定の青の群, D から F 判定の赤の群のカーネル特徴空間上の重心からそれぞれの点への距離を算出し, 実際の判定結果である群に含まれなかった誤った点がどのくらい存在するかを求めた. 特に, 今回用いたデータが健康診断データであるため, 赤の群に誤って判別されてしまった誤判別率に注目した. その結果を表 1 に記す.

表 1 に記す通り, 7 変数と変数を減らした 5 変数との誤判別率の差が 1 %未満という結果から AIC による結果から, 更に変数を減らしても大きな影響はないと考えられる.

4.4　SVM による基準値の推定

SVM(サポートベクターマシーン) とは, 2 クラスのパターン識別器を構成するひとつの手法である. 線形分離可能であることを前提に, データの中で, 最も他のクラスと近い位置にプロットされたものを基準とし, そのユークリッド距離が最も大きくなる (マージンが最大となる) ような位置の分離平面 (超平面) を求め, 識別器を構成する.

しかし, 複雑なデータでは, 線形分離可能であることは滅多にないため, ここでも, カーネル関数を利用する. データを高次元のベクトル空間に写像し, その空間で線形分離させることにより, 元のデータを非線形に分離することにする.

今回は, ある判定に有効とされた 2 成分における基準値を知るために SVM を利用する手法を考える.

図 4: SVM の結果 (T-Cho と TG)

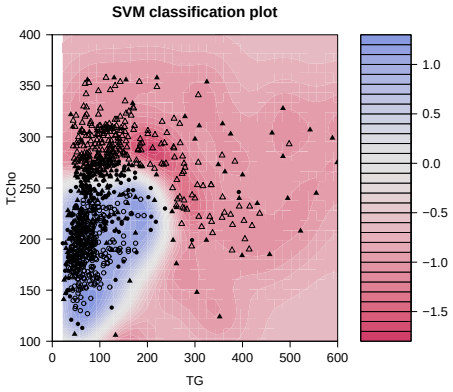


表 2: 基準値の比較

成分	推定 1	推定 2	基準値
T-Cho	100-250 mg/dl	100-250 mg/dl	140-219 mg/dl
HDL		40-100 mg/dl	40-96 mg/dl
TG	50-230 mg/dl		50-149 mg/dl

5　まとめ・課題

本研究の手法を用いることにより, ある評価データを説明する多変量データから有効な変量を絞り込み, その基準値まで求めることが可能であることを実データで実践した. 今後の課題として, 本研究の工程を自動化することが挙げられる.

参考文献

[1] 赤穂昭太郎,"カーネル多変量解析 非線形データ解析の新しい展開"(2008)

[2] 竹内友美,"カーネル PCA と SVM による高次元データの変数削減", お茶の水女子大学卒業研究会要旨集,pp69-70,(2011)