カーネル PCA と SVM による高次元データの変数削減

竹内友美　(指導教員：吉田裕亮)

1　はじめに

日頃, 私たちは様々な情報を利用して, 判断をしている. 与えられた多くの情報の中から, よりスムーズに判断を下すために, 判断に必要な情報だけを取り出すことが重要となる.

高次元データによる判別は, 計算量が多く時間がかかる. そこで, 変数選択を行う. 変数選択とは, 多くの変数の中から, 判別に必要な変数を選び出すことである. 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ある. 変数増加法は, 前進選択法とも呼ばれ, 判別に必要と考えられる変数を取り出す方法である. 変数減少法は, 後退消去法とも呼ばれ, 判別に必要ないと考えられる変数を取り除く方法である. 変数増減法は, ステップワイズ法とも呼ばれ, 判別に必要と考えられる変数を取り出し, その過程で判別に必要なくなったと考えられる変数を取り除く方法である. これらの方法により, 高次元データにおいて, 判別するための計算量を小さくすることができる.

そこで本研究では, 2 群に判別される高次元データにおいて, カーネル PCA と SVM を用いて, 誤判別率を求めることで, 変数減少法のように, 判別に必要でないと推定される変数をデータから除く手法のひとつを提案する.

2　カーネル PCA

2.1　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とら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, 線形な構造だけ見ていては不十分なことも多い.

2.2　カーネル法

カーネル法とは, カーネル関数を利用し, 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にし, 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. このことを, 一般にカーネルトリックという. すなわち, カーネル関数 K を用いて.

$$\langle \Phi(X), \Phi(Y) \rangle = \langle k(\cdot, x), k(\cdot, y) \rangle = K(x, y)$$

と考えることで, 実際に計算する際に, 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. よく使われるカーネル関数として以下のようなものがある.

線形カーネル

$$K(x, y) = x^T y.$$

(これは線形手法そのものである)

ガウスカーネル

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

多項式カーネル

$$K(x, y) = (c + x^T y)^d, \qquad (d : \text{自然数}, c \geq 0).$$

ラプラシアンカーネル

$$K(x, y) = \exp\left(-\beta \sum_{i=1}^{m} |x_i - y_i|\right), \quad (\beta > 0).$$

2.3　カーネル PCA

カーネル PCA とは, 2.2 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 を非線形化でき, 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, 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 2 主成分まで求めてプロットし, 適していると考えられるカーネル関数及びパラメータの値を見つけることにする.

3　SVM

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なる.

4　提案手法

4.1　元データの判別

2 群に判別される高次元データが与えられたとする. 平面で, 視覚的に観察することができるよう, 特徴空間を 2 次元に指定し, カーネル PCA を行う. 2 次元に縮約されプロットされたデータに対して SVM によりクラス分けし, 誤判別率を求める.

4.2　削減変数の選択

次に, 削減する変数の選択をしていく. もとのデータからある 1 変数を削減したデータを用意する. このデータを, 同様に特徴空間を 2 次元に指定し, カーネル

PCA をおこなう. 元のデータでの SVM によるクラス分類にあてはめ, 最も誤判別率の低い変数を削減の候補とする.

この削減手法を繰り返し, 元のデータによる SVM での誤判別率を超える 1 つ前までに選択された変数を削減してもよい変数と考え, 変数削減を行うことにする.

5 実データへの応用

年齢, 身長, 体重, 肥満度, 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）の 7 次元データのうち, 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）により, A から F に判定された 800 人分のデータを利用する. 判定が A から C の人, D から F の人の 2 群にわけ, 変数削減を行う. カーネル PCA, SVM 共に, ラプラシアンカーネルをカーネル関数に用いた. 元のデータによる SVM の結果は図 1 ようになる.

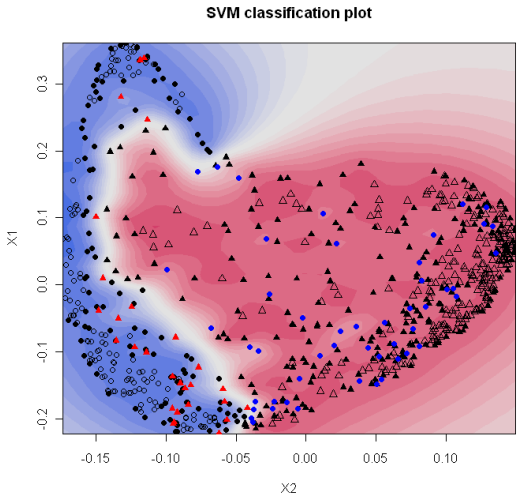

図 1: 全 7 変数の結果

このときの, 誤判別率は 9.75%となった.

次に, 元のデータから変数を削除する. 1 変数削減したときの誤判別率は, 以下の表のようになった.

変 数	誤判別率
年齢	0.09500
身長	0.09875
体重	0.09875
肥満度	0.09875
T-Cho	0.28750
HDL	0.09750
TG	0.54000

表 1: 1 変数削減後の誤判別率

このようになり, 誤判別率が 9.5%で最も小さい年齢の変数が削減候補となる.

次に, 年齢の変数を削減したデータから, 新たに 1 変数を削減. すると, 次は, 誤判別率が 9.5%の肥満度の変数が削減候補となる. 同様に, 変数削減していくと, 次に体重の変数, その次に身長の変数が誤判別率最少で, 削減候補となる.

年齢, 身長, 体重, 肥満度の変数を削減したのち, もう 1 変数削減しようとしたときの, 誤判別率は, 以下の表のようになった.

変 数	誤判別率
年齢	0.09750
身長	0.09750
体重	0.09750
肥満度	0.09750
T-Cho	0.29500
HDL	0.10750
TG	0.51250

表 2: 4 変数削減後の誤判別率

どの変数を削減しても, 元のデータでの SVM による結果の誤判別率 9.75%を超えてしまうので, もう変数削減できないもとの考える.

そこで, 年齢, 身長, 体重, 肥満度の変数を削減したデータで, プロットしてみた結果が図 2 である.

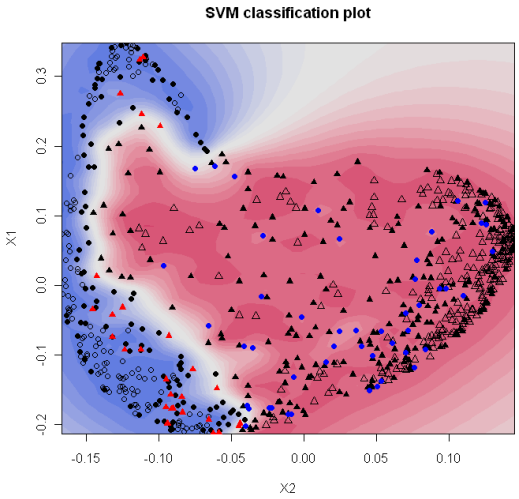

図 2: 4 変数削減後の結果

このようになり, 元のデータ全 7 変数でプロットした図 1 と, 大きな差がないことがわかる.

6 まとめ・課題

カーネル PCA と SVM による変数削減は有効であると考えられる. しかし, この提案手法では, どのカーネルを利用するかによって結果が大きく変わる場合もあり, どのカーネルが適していてるかはアドホックな手法である点の改良は今後の課題である.

参考文献

カーネル法　正定値カーネルを用いたデータ解析
http://www.ism.ac.jp/~fukumizu/ISM_lecture_2004/Lecture2004_kernel_method.pdf
Support Vector Machine
http://www.neuro.sfc.keio.ac.jp/~masato/study/SVM/index.htm