社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－MPS－73 (16)
2009／3／5

バギングを用いた2次元非線形判別曲線の推定

白川 聖子* 吉田 裕亮**
*お茶の水女子大学 理学部 情報科学科
**お茶の水女子大学大学院人間文化創成科学研究科

2次元データの非線形2群判別における, 判別曲線の推定を行う手法として, ニューラルネットワーク法などがある. しかしこれらの手法は推定すべきパラメータが多く必要であり, 一般に計算が繁雑である. 本研究では, 比較的簡単な推定を重ね合わせるバギング法を用いた非線形判別曲線の推定手法, 並びに幾つかの数値実験結果について報告する.

Estimation of two dimensional nonlinear discriminant curves by bagging method

Seiko Shirakawa*, Hiroaki Yoshida**
*Department of Information Sciences, Ochanomizu University
**Graduate school of Humanities and Sciences, Ochanomizu University

Neural network is known as one of techniques for estimation of two dimensional nonlinear discriminant curves. In this technique, however, it should be required to control many parameters, and the calculation is complicated in general. In this study, we use bagging method in order to estimate the discriminant curves, in which using bootstrap samples from the training sets and then we aggregate to form a bagged discriminant, and report some numerical experiments.

1 はじめに

ニューラルネットワーク法などによる非線形判別は線形判別に比べて, 一般に多くのパラメータの推定が必要であり, 計算が繁雑である. そこで現実にはより簡単な方法で非線形な判別曲線を推定することが求められる. 本研究ではこのような方法のひとつとして, バギング法を用いた非線形判別曲線の推定手法を提案する. バギングにおける1つ1つの仮説は, 弱仮説と呼ばれるものであり比較的簡単な推定であるが, それを重ね合わせることにより最終的に強い仮説を推定する.

2 バギング法

バギング (bagging) とは bootstrap aggregating に由来し, その名のとおりブートストラップ法により例題をリサンプリングして異なる弱い仮説を多数作り, それらから集合体を構成することによって最終的な仮説を作る方法一般を指す. なお仮説の生成は並列的であり, リサンプリングは独立に行うので, 弱仮説どうしは互いに影響しない.

バギングによる仮説の構成方法を説明する. N 個の例題からなる訓練集合が与えられているとする.

ステップ 1: 例題より m 回復元抽出し例題を集め, これを用いて仮説 h を学習する.

ステップ 2: ステップ 1 を B 回行い, 仮説を B 個 $\{h(x;\theta_i); i = 1, \cdots, B\}$ 構成する.

ステップ 3: 回帰問題の場合には

$$H(x) = \frac{1}{B}\sum_{i=1}^{B} h(x;\theta_i)$$

により, 判別問題では

$$H(x) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} |\{i|h(x;\theta_i) = y\}|$$

$$= \underset{y\in\mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{B} I(h(x;\theta_i) = y)$$

により, 最終的な仮説を構成する.

3 非線形判別曲線の推定法

2次元データの非線形2群判別にバギング法を用いた推定を行うため, 以下のような操作を施す.

3.1 推定のアルゴリズム

1. データ領域内にランダム領域をとる.
2. ランダム領域内からデータをいくつか抽出し, それぞれ0値, 1値を取るものに判別する.
3. 0値を取るもの, 1値を取るもの各群の重心座標を求め, ランダム領域内で2点に関するマハラノビス距離による中点の軌跡を求める.
4. 工程1～3をN回繰り返し, データ領域内においてn回以上通った点を再度プロットする.
   N: 指定する十分大きな自然数
   n: Nに応じて適宜指定される自然数

3.2 ランダム領域

上のアルゴリズムにおいて, より精密な判別曲線を浮かび上がらせるためには, データ領域内において, できる限り偏りの少ないランダム領域を取ることが必要となる.

ここで本研究ではデータ領域よりも大きな, 架空領域を考える. 架空領域内で乱数s_1, t_1, uを与え(s_1, t_1)座標を1点, uを一辺の長さとする正方形をランダム領域とする.

この手法は簡単な領域の取り方である. しかしデータ領域内だけで行うと偏った部分に多く領域が現れる. 例えばs_1, t_1にuを足したものを他の頂点とすると考えると, データ領域内における右上の部分に多く領域が現れる. そこで架空領域を考える. するとランダム領域は, 架空領域内では偏りがあるものの, データ領域内での偏りはかなり減少させることができる.

3.3 マハラノビス距離

2次元データの分散共分散に基づいて定義される距離がマハラノビス距離であり, 単純なユークリッド距離よりデータの特性が反映された距離である.

2変数データによるマハラノビス距離は以下のように与えられる. k個の群の各平均を$\vec{\mu_j} = (\mu_{1j}, \mu_{2j})^T$, $(j = 1, 2, \cdots, k)$, 観測値を$\vec{\mathrm{X}} = (X_1, X_2)^T$とする. 各群の分散共分散行列を$\sum_j$, その逆行列を$\sum_j^{-1}$とする. このとき第$j$群のマハラノビス距離$d_j$は

$$d_j = \sqrt{(\vec{\mathrm{X}} - \vec{\mu_j})^T \sum_j^{-1} (\vec{\mathrm{X}} - \vec{\mu_j})}$$

で与えられる. したがって,

$$\sum_j^{-1} = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}$$

とすると, ある群の平均(x_0, y_0)から(ξ, η)までのマハラノビス距離は

$$(\xi - x_0)^2\alpha + 2\gamma(\xi - x_0)(\eta - y_0) + (\eta - y_0)^2\beta$$

となる. このことより0値群と1値群, 各群の重心からマハラノビス距離による等距離点を描くと, 2次曲線が得られる. ユークリッド距離に基づく中点を取るよりも, より各群の分散や相関が取り込まれた推定曲線が現れる.

3.4 シミュレーションデータでの実験

2次元の有界領域$D = [-2, 2] \times [-2, 2]$内で与えた曲線により, それぞれ0値, 1値が割り当てられたデータを計1000個用意する. 各データは[x座標, y座標, $\{0, 1\}$]の3つの値が与えられている. 以下のシミュレーションは$N = 8000, n = 5$と設定されている.

群の設定： $y \geq x^3 - x \rightarrow \{0\}$
$y < x^3 - x \rightarrow \{1\}$

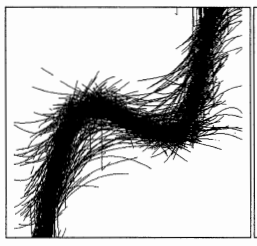


図 1: N=8000

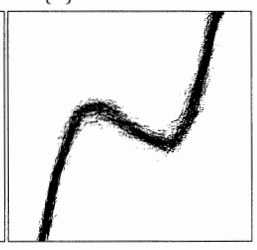


図 2: n=5

群の設定： $x^2+y^2 \leq 1 \to \{0\}$
$x^2+y^2 > 1 \to \{1\}$

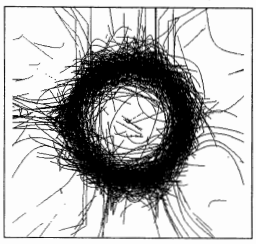


図 3: N=8000

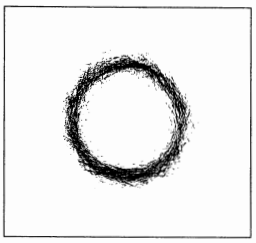


図 4: n=5

図 1,3 はアルゴリズムによって描かれた, 8000 本の曲線である. 図 2,4 はそれぞれ図 1,3 において, 5 回以上通った点のみを取り出したものである.

図 2,4 を可能な限り 1 本の曲線に近づけるため, 本研究では以下の 2 段階の画像補正を施した.

4 画像補正

4.1 補正アルゴリズム 1

1. 画像を pbm データとして保存する.

2. ある点 a を中心とする $9(=3\times3)$ 点のうち 1 が m 点以上の場合 $a=1$, m 点未満の場合 $a=0$ とする. (m: 9 以下の指定する自然数)
なお 9 点を $25(=5\times5)$ 点と変えて, 大きな範囲を基に補正を行うこともできる.

3. 工程 2 の m を変えながら M 回繰り返す.
M: 指定する自然数 (10 ~ 20 程度)

すなわちこの補正アルゴリズム 1 によって, 画像を 1 本の太い曲線に近づけることになる. 次に太い 1 本の曲線を細くする, 以下の補正アルゴリズム 2 を施す.

4.2 補正アルゴリズム 2

1. 画像を pbm データとして保存する. 順に補正を行うが, 各点の元の値を α とし, 補正後の値を β とする. $(\alpha,\beta \in \{0,1\})$

2. 各行 (横方向), 両端については $\beta=\alpha$.

3. その他の点については
(1) $\alpha=0$ のとき,
(i) 並びが $1\alpha1$ ならば $\beta=1$,
(ii) それ以外ならば $\beta=0$,
とする.
(2) $\alpha=1$ のとき,
(i) $0\alpha0, 1\alpha1$ ならば $\beta=1$,
(ii) $0\alpha11\cdots1, 1\cdots11\alpha0$ ならば,

$$\begin{cases} 1 \text{の個数が } L \text{ 個より大のとき} \beta=0, \\ 1 \text{の個数が } L \text{ 個以下のとき} \beta=1, \end{cases}$$

とする.

4. すべての行に対して工程 2 ~ 3 を行う.

5. 工程 2 ~ 4 を列 (縦方向) で行う.

6. 横方向, 縦方向の補正を交互に複数回繰り返す.

L の値によって最終的な曲線は定まる. また L の値に依存するが, 横縦方向を 1 セットと考え, 約 30 セット以上繰り返すと, 画像はほぼ一定の状態に落ち着く.

4.3 画像補正を施した例

推定法のシミュレーションで使用したデータの結果を使用する. なお $L=5$ とし, 30 セットの繰り返しを行った.

群の設定： $y \geq x^3-x \to \{0\}$
$y < x^3-x \to \{1\}$

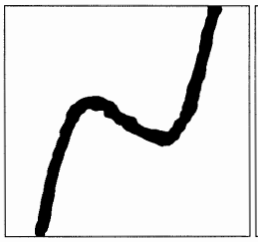


図 5: 補正 1 後

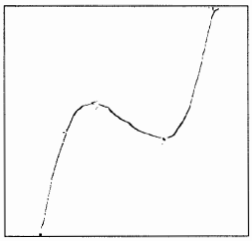


図 6: 補正 2 後

群の設定： $x^2+y^2 \leq 1 \to \{0\}$
$x^2+y^2 > 1 \to \{1\}$

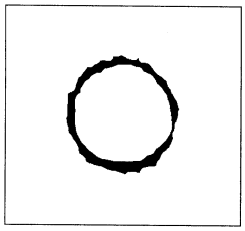


図 7: 補正 1 後

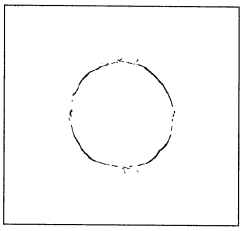


図 8: 補正 2 後

図 5,7 は補正 1 を m の値を変えながら, 15 回行ったものである. 図 6,8 は, 図 5,7 に補正アルゴリズム 2 を適用した結果で, $L=5$ であることから縦, 横共に 5 ピクセル以下となる.

5 実データへの応用

実データへの応用として, 最高血圧値, 最低血圧値に基づいて血圧疾患の判定が行われた健康診断のデータを用いた. 正常 (0 値), 異常 (1 値), それぞれ 130 人ずつとなるように抽出した.

以下横軸が最高血圧値, 縦軸が最低血圧値である.

- 推定のアルゴリズムを施した結果

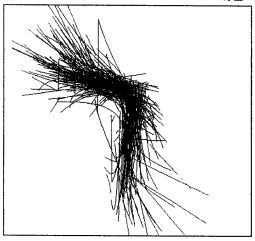


図 9: $N=8000$

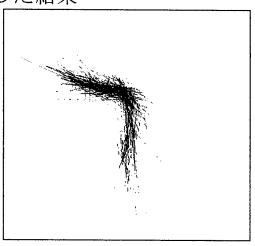


図 10: $n=3$

- 補正アルゴリズム 1,2 を施した結果

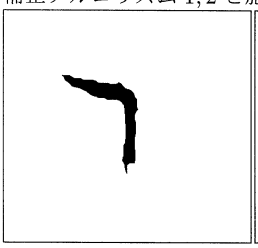


図 11: 補正 1 後

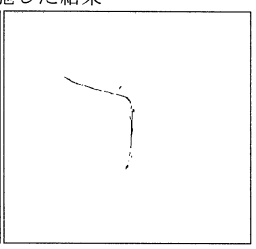
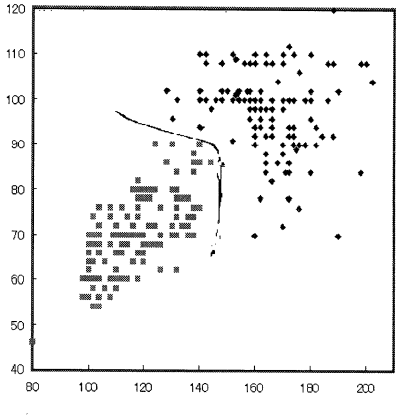


図 12: 補正 2 後

- データと図 12 を重ね合わせた結果



WHO の血圧判定基準によると, 最高血圧値 140 mmHg, 最低血圧値 90mmHg からが高血圧とみなされる. 上の結果より, 用いたデータはこの基準に基づいて血圧異常の判定を行っていると推定できる.

6 まとめ

バギング法と簡単な画像処理を用いた, 2 次元非線形判別曲線を推定するひとつの手法を提案した.

シミュレーションデータの場合, 事前に設定したデータ値の境界線とほぼ同じ曲線 (特に円の場合においても) が推定可能であった. 実データへの応用では, シミュレーションと比べデータ数が少なく, データのばらつきに偏りがあるにも関わらず, ある程度の判別曲線の推定が得られた. したがって, 本研究で提案した手法は 2 次元データの非線形な判別曲線の推定に有効な手法のひとつであると思われる. 今後の課題として, 小数データに関する適応法なども検討したい.

参考文献

1. 麻生英樹, 津田宏治, 村田昇, パターン認識と学習の統計学〜新しい概念と手法〜, 岩波書店 (2003).