日本オペレーションズ・リサーチ学会
2004年春季研究発表会

ブースティングを用いたスコアリングモデルの構築

02302970 筑波大学 竹林 実 TAKEBAYASHI Minoru
02203190 筑波大学 佐野夏樹 SANO Natsuki
01207840 筑波大学 *鈴木秀男 SUZUKI Hideo

1. はじめに

近年データウェアハウスやデータベースが普及したことから，大量のデータが蓄積されるようになっている．それに伴い，大量のデータから有効な情報を抽出することのできる手法が注目されている．その一つにスコアリングというものがある．スコアリングとは，既存のデータを基に，ある顧客について予測される事象が起こる可能性を推定するものであり，幅広い分野に応用されており（例えば文献 [3]），その精度の向上が望まれている．

一方，2クラスの分類問題の予測精度を向上させるために用いられる手法にブースティングがある．スコアリングは可能性といった連続値を予測する問題であるが，本質的には2クラスの分類問題と同様の問題であると考えることができる．よってブースティングをスコアリングへ応用することで，その精度の向上が期待される．代表的なブースティング手法の一つに AdaBoost[1] があり，竹林，佐野，鈴木 [4] は AdaBoost を用いて顧客スコアリングモデルを構築している．本研究では，AdaBoost と同様の枠組みで，異なった損失関数を扱うことのできる MarginBoost 及び MarginBoost.L1[2] の2つのブースティング手法を用いてスコアリングを行うことを提案し，従来法と比較することでその有効性を検証する．

2. ブースティング

ブースティングは，精度の低い学習機械（基本学習機械）を組み合わせることで，精度の高い学習機械を構成する手法である．本研究で用いる MarginBoost 及び MarginBoost.L1[2] のそれぞれのアルゴリズムをまとめたものを以下の図1に示す．

ここで学習データは $S = \{(x_i, y_i) : i = 1, \ldots, N\}$ であり，$y \in \{1, -1\}$ の2クラスの分類問題を考える．図中のステップ0では損失関数を決める．ここで指数関数を用いた場合，MarginBoost は AdaBoost と同等の手法となる．ステップ1では初期化を行う．ステップ2 (a), (b) では重み付きリサンプリングを行うことで，分類の難しいデータもうまく分類できる基本学習機械 f_t を生成している．ステップ2 (c) の信頼度 β_t については，本研究では損失関数を最小とする β をラインサーチにより求めた．ステップ2 (d) の学習機械の構成方法が MarginBoost と

0. 適切な損失関数 C を決める．
1. 学習データの重みを $w_1(i) = 1/N$，学習機械を $F_0(\mathbf{x}_i) = 0$ と初期化する． $i = 1, \ldots, N$
2. t=1, ... ,T に対し，
(a) 重み w_t により学習データ S のリサンプリングを行い，それを $\tilde{S}_t$ とする．
(b) $\tilde{S}_t$ を用いて学習し，基本学習機械 f_t を生成する．
(c) 適切な信頼度 β_t を選ぶ．

MarginBoost は (d1)，MarginBoost.L1 は (d2) を行う．
(d1) 学習機械を $F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \beta_t f_t(\mathbf{x})$ とする．
(d2) 学習機械を $F_t(\mathbf{x}) = \frac{\sum_{m=1}^{t} \beta_m f_m(\mathbf{x})}{\sum_{m=1}^{t} \beta_m}$ とする．
(e) 学習データの重みを以下のように更新する．
$$w_{t+1}(i) = \frac{C'(y_i F_t(\mathbf{x}_i))}{\sum_{j=1}^{N} C'(y_j F_t(\mathbf{x}_j))}, \quad i = 1, \ldots, N$$
3. 最終学習機械として $\mathrm{sign}(F_T(\mathbf{x}))$ を得る．

図1: MarginBoost 及び MarginBoost.L1 アルゴリズム

MarginBoost.L1 の相違点である．ステップ2 (e) では，f_t において分類の正解したサンプル (x_i, y_i) に対しては，次のラウンドの重み $w_{t+1}(i)$ を小さくし，誤ったサンプルに対しては $w_{t+1}(i)$ を大きくするという考えに基づいた重みの更新が行われている．そして，ステップ3で T 個の基本学習機械の重み付き結合により，1つの学習機械へ統合し，その符号をとったものを，1または -1 を出力する最終的な学習機械としている．

3. 利用データ

本研究で利用したデータは先行研究 [3], [4] で利用されている，ある衣料・雑貨販売会社の通信販売履歴データと，UCI Machine Learning Repository[5] に公開されている German Credit データの2種類のデータである．これらのデータをそれぞれ，スコアリングモデル構築に用いる学習データと，モデルの性能検証に用いるテストデータとに二分して用いる．ここでは通信販売履歴データの分析結果のみを報告し，German Credit データの分析結果については，発表の際に報告する．

通信販売履歴データは，取引 ID をキーに持つ販売履歴データに，商品属性・顧客属性に関する情報を付加したものである．このデータには約3年間分の販売履歴が記録されている．これをもとに顧客 ID をキーとした分析用データを作成した．前半 30ヶ月を入力期間として

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

表 1: 分析用データ

属性	データの内容
key	[顧客ＩＤ]
in1	入力期間中の購入金額合計
in2	入力期間中の購入回数合計
in3	入力期間中の最新の購入日までの日数
in4	入力期間中の 2 番目に新しい購入日までの日数
in5	入力期間中の 3 番目に新しい購入日までの日数
out	予測期間中の購入有無

説明変数の作成に使用し，後半 4ヶ月を予測期間としてこの期間の商品の購入有無を目的変数（クラス）とし，入力期間に 1 回以上の取引があった顧客 10,560 名について分析用データを作成した．分析用データに用いた変数を表 1 に示す．また予測期間に取引があった顧客（購入者）は 1,032 名で，全体の約 10%にあたる．

4. スコアリングモデルの構築

本研究では，図 1 中の $F_T(\mathbf{x})$ をスコアリングモデルと考える．それにより出力される連続値を，可能性を示すスコアとする．スコアリングモデルの構築の際には，基本学習機械とブースティング手法の組み合わせにより 40 通りのモデルを構築し，最もスコアリングの精度の高いものを本提案モデルとして採用した．基本学習機械は，深さが 1 から 4 の 4 通りの決定木（分類木）を用いた．ブースティング手法は，MarginBoost 及び MarginBoost.L1 に対して，それぞれ 5 通りの損失関数を用いたものの計 10 通りを用いた．またブースティングのアルゴリズム中で行われるリサンプリングの誤差を考慮し，本研究では各組み合わせに対し 5 回ずつモデルを構築し，その平均値を各モデルによる結果として用いた．

スコアリングモデルの評価には，累積ゲイン図とリフト率を用いる．累積ゲイン図は，予測全体の有効性を評価するもので，スコアの高い順に顧客一覧を並べ替えたときに，予測上位 x%の顧客について反応者数の累積割合をプロットしたものである．リフト率は，予測スコア上位 x%の顧客を抽出した場合のモデルのあてはまりの良さを定量的に評価するもので，以下の式で定義される．

$$\text{リフト率}\,(x) = \frac{\text{予測上位 } x\%\text{の顧客の反応率}}{\text{全顧客についての反応率}}$$

5. 結果

本提案モデルと，回帰木（RT），ロジスティック回帰分析（LR）の 3 つの手法によるスコアリング精度の比較を行った．本提案モデルは，深さ 1 の決定木を基本学習機械として，指数関数を損失関数として用いた MarginBoost.L1 を適用したものである．

累積ゲイン図（図 2）より，本提案モデルは，決定木よりも精度が高く，またロジスティック回帰分析とほぼ同等の精度が得られたことが分かった．表 2 のリフト率により定量的に比較をすると，予測スコア上位 20%までは本提案モデルが最も良い値が得られており，上位 30%以降においても，ロジスティック回帰分析によるモデルとほぼ同等の値が得られている．つまり本提案モデルはスコア上位者の予測に優れており，優良顧客の選出する場合などにおいて，非常に有効であることが示された．

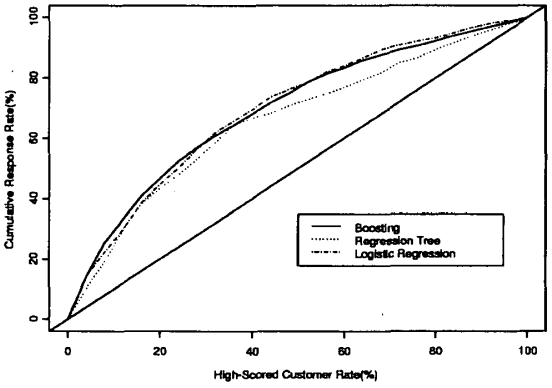

図 2: 累積ゲイン図

表 2: 予測スコア上位 10%毎のリフト率の比較

手法	10%	20%	30%	40%	50%
Boost	2.98	2.33	1.96	1.70	1.54
RT	2.41	2.17	1.86	1.67	1.43
LR	2.66	2.23	1.97	1.74	1.55

手法	60%	70%	80%	90%	100%
Boost	1.39	1.27	1.15	1.07	1.00
RT	1.28	1.19	1.12	1.05	1.00
LR	1.40	1.29	1.17	1.08	1.00

"Boost"は本提案モデルを表す．

参考文献

[1] Freund, Y. and Schapire, R.E. "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Comp. and System Sci.*, **55**(1), 119-139. (1997)

[2] Mason, L., Baxter, J., Bartlett, P.L. and Frean, M. "Functional Gradient Techniques for ComBining Hypotheses", In Smola, A.J., Bartlett, P., Scholkopf, B., and Schuurmans, C. (Eds.), *Advances in Large Margin Classifiers.* MIT Press, Cambridge, MA. (2000)

[3] 後藤正輝，村山一穂，門間公志，香田正人「データマイニング手法によるスコアリングモデルの開発」，Direct Marketing Review, vol.1, 19-32. (2002)

[4] 竹林実，佐野夏樹，鈴木秀男「AdaBoost による顧客スコアリング」，2003 年日本オペレーションズ・リサーチ学会秋季研究発表会アブストラクト集, 288-289. (2003)

[5] UCI Machine Learning Repository. http://www.ics.uci.edu/ mlearn/MLRepository

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．