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

AdaBoostによる顧客スコアリング

（申請中）　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

顧客スコアリングとは，過去の購買履歴データを基にして，購入可能性の高い順に顧客をランク付けする問題である．スコアリングモデルの構築は，ある期間に顧客が商品を購入するか否かを，それ以前の購買行動から予測するモデルを作成することで行われる．予測値として，購入するか否かの二値ではなく，連続値を得ることで，その値を基に顧客をランク付けする．顧客スコアリングは，データマイニング手法の応用分野としても注目されている．(例えば文献 [2])

本研究では，学習機械として決定木を用いたAdaBoostにより，現実の企業の取引履歴データに対する顧客スコアリングモデルを作成する．また従来の手法と比較することで，その有効性を検証する．

2　AdaBoost

Boostingは精度の低い学習機械(Simple Learner)を組み合わせることで，精度の良い学習機械を構成する手法である．その代表的なものとして，多くの理論的な検証と実験的実証がなされてきたアルゴリズムに「AdaBoost」[1] が挙げられる．以下にそのアルゴリズムを示す．

AdaBoost

For $t = 1$ to T

$\widetilde{S}_t = \mathrm{ReSam}(S, D_t)$

$f_t = M(\widetilde{S}_t)$

$\epsilon_t = \Pr_{i \sim D_t}\{f_t(x_i) \neq y_i\}$

$c_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$

$$D_{t+1}(i) \propto D_t(i) \times \begin{cases} e^{-c_t} & \text{if } f_t(x_i) = y_i \\ e^{c_t} & \text{if } f_t(x_i) \neq y_i \end{cases}$$

EndFor

Combined Learner: $G(x) = \sum_{t=1}^{T} c_t f_t(x)$

ここで学習データは $S = \{(x_i, y_i) : i = 1, \ldots, n\}$ であり，$y \in \{1, -1\}$ の二値判別問題を考える．またサンプル S を確率 D でリサンプリングすることを，ここでは $\mathrm{ReSam}(S, D)$ と表現する．

ラウンド t において，$\mathrm{ReSam}(S, D_t)$ により重み付きリサンプリングを行い $\widetilde{S}_t$ を受け取って，学習機械 M により Simple Learner f_t を出力する．続いて f_t の誤り率 ϵ_t と，f_t の信頼度 c_t を求める．ここで f_t において正解したサンプル (x_i, y_i) に対しては，次のラウンドの重み $D_{t+1}(i)$ を小さくし，一方誤ったサンプル (x_i, y_i) に対しては $D_{t+1}(i)$ を大きくする．つまり判別の難しいサンプルを重点的に学習していく．これらを T ラウンド繰り返す．最終的な判別は，T 個の Simple Leaner f_t の信頼度 c_t による加重和 (Combined Leaner) の符号をとることで行われる．

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

3　利用データ

本研究で利用したデータは先行研究 [2] 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である．原データは取引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．この原データをもとに顧客ＩＤをキーとした分析用データを作成した．

原データには約３年間分の販売履歴が記録されており，前半30ヶ月を入力期間として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応答変数とした．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．

4 スコアリングモデルの作成

分析用データを学習データとテストデータに二分し，学習データによりスコアリングモデルを作成し，テストデータによりモデルの汎化能力の検証を行う．ここで学習データのサンプル数は1,056，テストデータのサンプル数は9,504とした．またテストデータにおける反応者数は920人であり，全顧客の反応率は9.7％であった．

本研究では，学習機械として深さ3の決定木を用いたAdaBoostを適用した．顧客スコアリングモデルとして，Combined Leaner $G(x)$ を用い，その出力値を顧客スコアとする．出力値が大きいほど，購入可能性の高い優良顧客であると判断される．

$$\text{顧客スコアリングモデル}：G(x) = \sum_{t=1}^{T} c_t f_t(x)$$

スコアリングモデルの精度評価の基準として，予測全体の有効性を評価するものとして累積ゲイン図 [3] を用いた．累積ゲイン図はモデルに基づき，顧客をスコアの高い順にソートしたときの反応者数の累積割合をプロットしたものである．また，予測スコア上位 x ％の顧客群に対する精度評価には次のリフト率を用いた．

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

5 結果

学習機械として決定木を用いたAdaBoostと，決定木，ロジスティック回帰の3つの手法で予測精度の比較を行った．AdaBoostにおけるラウンド数は，多くしすぎると学習データに対するオーバーフィッティングが見られたため，本研究では50とした．また決定木は，AdaBoostの適用にあたっては二値の反応変量を返す分類木として用い，単独で用いる場合は連続値を返す回帰木として用いた．

得られたモデルから累積ゲイン図（図1）を作成した．AdaBoostを適用したモデルは，決定木よりも予測精度が高く，またロジスティック回帰とほぼ同等の予測精度が得られた．予測スコアが上位の顧客群に対するリフト率（表2）を見ると，上位10％（950人）の顧客についてはAdaBoostが最も良いが，それ以降はロジスティック回帰が若干よくなっている．しかしスコアの最も高い優良顧客の選出することを考えた場合は，上位10％の顧客についてのリフト率が最も高いということは現実的に有効であり，AdaBoostの有効性が示せたと言える．

表2: リフト率の比較

手法	10 %	20 %	30 %	40 %	50 %
AdaBoost	2.84	2.20	1.83	1.67	1.50
DT	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 %
AdaBoost	1.37	1.26	1.17	1.08	1.00
DT	1.28	1.19	1.12	1.05	1.00
LR	1.40	1.29	1.17	1.08	1.00

"DT"は決定木，"LR"はロジスティック回帰

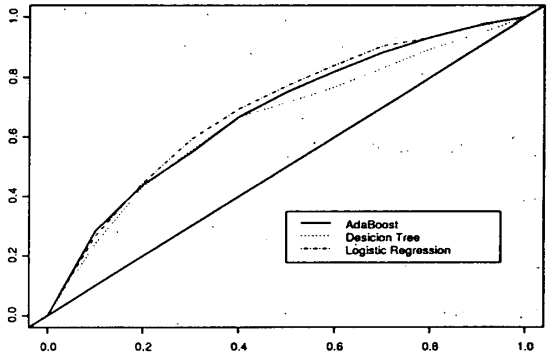

図1: 累積ゲイン図．横軸はスコアの高い順にソートされた顧客の割合，縦軸は累積反応者割合

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**, 119-139. (1997)

[2] 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)

[3] マイケルJ．A．ベリー，ゴードン・リノフ 著，江原 淳，金子 武久，斎藤 史朗，佐藤 栄作，清水 聰，寺田 英治，守口 剛訳『マスタリング・データマイニング ＜理論編＞』（海文堂，2002）