2000年度日本オペレーションズ・リサーチ学会
秋季研究発表会

決定木を用いた複合学習モデルについて

筑波大学大学院経営政策科学研究科　*山部浩司　YAMABE Hiroshi
(株) ダイエーオーエムシー　八巻　智　YAMAKI Satoshi
(株) ダイエーオーエムシー　山本良次　YAMAMOTO Yoshitsugu
01105930　筑波大学　香田正人　KODA Masato

1　はじめに

データマイニングにおける予測モデルには, 属性値に対する解釈や学習データを柔軟に選択することが要求される. 予測モデルとして決定木 (回帰木) が多用されるが, 学習データの偏りにより誤判別が発生することは避けられない. テキストマイニングでは, 適応リサンプリング法を用いることにより決定木の精度を向上させている.

本論文では, 複数の決定木モデルに適応リサンプリング法を応用することで, 学習精度の向上を可能とする複合学習モデルを提案し, 数値実験を行いその妥当性について検証を行う.

2　データ

今回使用したデータは, ダイエーオーエムシーにおける顧客データの中から 1998 年 10 月に入会した分を使用した. この中で毎月のデータを 3ヶ月目から 12ヶ月目までの 10ヶ月間の履歴データを用いて分析を行った. 該当する顧客数は 16382 件である.

今回の分析では, カード利用の属性値の中でキャッシング利用に注目した. 各顧客が毎月使用したキャッシングの金額と件数を用いる. なお, 顧客セグメントは, 利用実績 (履歴) に基づき, あらかじめニューラルネットによるクラスタリングで 4 分類されている.4 クラスターは以下のとおりである.

cluster1：未使用者
cluster2：利用者 A(年度末, ボーナス期利用)
cluster3：利用者 B(年度末, ボーナス期未使用)
cluster4：高額利用継続者

利用パターンを分析する上で, ボーナス時の影響や利用者の比率の変動を考慮し, 比較的利用状況に特別な要因の発生していない 10 月入会の顧客に限定している.

3　複合学習モデル

今回の分析で使用した手法は図 1 のような, 決定木を用いた複合学習モデルである.

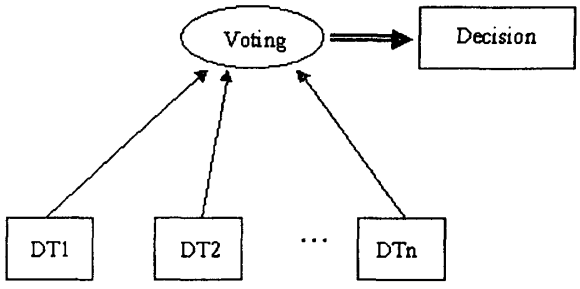

図 1: 複合学習モデル

複合学習モデルとは, 異なる学習データを用いた複数の決定木 (Clasification Tree) による結果を統合して最終評価を行う. 誤判別に注目した適応リサンプリングにより, 学習データを再構成することで, 既存の決定木分析よりも精度が上がるものと期待される.

全体のデータは 16382 件であり, これを学習データ 10000 件と試験データ 6382 件に分割した. 決定木は 3 個 (n=3) 作成した.

3.1　決定木モデル

第一の決定木 (以下「DT1」) は, オリジナルの学習データを利用し, 決定木を作成したものである.[1] この結果を基に, 分類結果のクラスターと, 真のクラスターとを比較する事で, 誤判別数を求める. クラスター j における擬似確率関数 pr(j) を以下で定義する.

$$pr(j) = \frac{1 + e(j)^m}{\sum_{i=1}^{n}(1 + e(i)^m)}$$

j=1,2,3,4
m:任意の正整数

e(j):クラスター j の誤判別数

この擬似確率を利用して第二の決定木 (以下「DT2」) で利用する学習データの比率を決定する. リサンプリングには, 復元を許した無作為抽出 [2] を行い,DT1 と同様 10000 件のデータを作り出す.DT1 での誤判別の高いク

ラスター程,DT2 における学習データ全体に占める割合が高くなる. これは, 誤判別の高いクラスターを学習データとして増加させることで, 学習効果を高めることが目的である.[3]

この DT2 モデルと DT1 モデルの分類結果を単純比較し, 異なる結果を導き出したデータのみを抽出して, 第三の決定木 (以下「DT3」) 作成用の学習データを構成する.DT1 と DT2 が同じ結果を導き出したデータを除外したのは, 投票によって採用される過半数に影響を与えず, 学習の意味が無いためである.

以上の方法でモデルを構築した後, 試験データを使用した分類結果で投票を行い, 過半数を得た結果をこの複合モデルによる最終予測 (以下「vote」) であるとする. 予測結果から真の値との誤判別率を求め,DT1 のみの結果と比較を行う.

4 数値実験

4.1 学習データ

学習データに基づく各クラスターの誤判別率を表 1 に示す. 誤判別率の下の括弧は, データの数である.DT1 では, クラスター 1 と 4 の誤判別率が低い. 逆に, クラスター 2,3 は,DT2 の誤判別率が低い. これらの結果から, 擬似確率を利用した決定木では, オリジナルな決定木で分類結果の悪かったデータに対して誤判別率を改善するという結果が得られた.

	DT1	DT2	DT3
cluster1	0.00	1.00	0.00
	(8420)	(1)	(8420)
cluster2	0.22	0.02	0.31
	(482)	(6541)	(111)
cluster3	0.19	0.02	0.68
	(371)	(3049)	(69)
cluster4	0.04	0.50	0.07
	(727)	(409)	(393)

表 1: 誤判別率 (学習データ)

4.2 試験データ

各決定木の誤判別数, 誤判別率と最終投票結果を表 2 に示す. 不定数とは, 各決定木においての結果が異なり, 投票結果が出なかったものである.

各決定木における誤判別率の傾向については学習データにおける結果と類似している. このことから, 教師付き学習における問題点である, 過剰学習 (Over Fitting) の可能性が低いモデルとなっている.

	データ	DT1	DT2	DT3	Vote	不定数
cluster1	5373	0.00	1.00	0.00	0.00	0
cluster2	323	0.21	0.03	0.54	0.08	3
cluster3	211	0.17	0.05	0.94	0.13	10
cluster4	475	0.07	0.61	0.06	0.11	14

表 2: 誤判別率と不定数 (試験データ)

投票結果と DT1 を比較すると, クラスター 4 を除く各クラスターで DT1 のみの予測に比べ誤判別率の改善が見られる.

不定数の数の全体に占める割合は 1%以下であり, 最も割合の高いクラスターでも 5%以下である.

5 おわりに

今回, 適応リサンプリング法に基づく複合学習モデルにより分類精度が向上し, その有効性を数値実験により検証できた.

今後の課題としては, 決定木をさらに増やした場合の精度や, 他のデータマイニング手法と複合的に組み合わせたモデルとの精度比較を行っていく予定である.

参考文献

[1] J.M. チェンバース,T.J. ヘイスティ編, 柴田里程訳『S と統計モデル』(共立出版 1994 年)

[2] B.Efron & R.J.Tibshirani, An Introduction to the Bootstrap, New York, Chapman&Hall, 1993

[3] G.Dupret & M.Koda, "Bootstrap Re-Sampling and Cross-Valiation for Neural Network Learning," Discussion Paper Series No.853 Inst. Policy and Planning Sciences, University of Tsukuba, March 2000 (forthcoming, European Journal of Operational Research)