総合通販企業におけるアンサンブルアルゴリズムを用いた顧客の取引継続に関する研究

森田　裕之

キーワード：データマイニング，分類問題，アンサンブルアルゴリズム

本稿は，西口 浩司さんの 2015 年度大阪府立大学大学院経済学研究科に提出された修士論文をもとに，加筆修正したものです．

1. 問題の説明と得られた結果

近年，B2C 市場への参入者が増大し，その取引の種類も多岐にわたるにつれ，これまでカタログ販売を主たる販売チャネルとしていた総合通信販売企業にとって，強力な競合企業が出現することになり，その競争はますます激化している現状にあります．

幸いなことに，取引の性質上，ID 付 POS データは長期間にわたって取得されているため，データマイニングの観点から，当該企業にとって必要な情報や知識をマイニングし，企業経営の意思決定を支援する材料として，提供することが可能です．データマイニングの適用方法は，さまざま考えられますが，ここでは，CRM の観点から顧客の生涯取引価値に焦点を当て，顧客が取引を中止する原因や，逆に，優良顧客が取引を増大する要因を発見するために，まずは全体のデータから適切な顧客セグメントを識別して，分類問題を定義します．定義された分類問題に対して，可読性が高く，またモデルの分類精度も頑強であることが期待される分類モデルとして，ここでは複数の分類モデルを組合せたアンサンブルモデルを提案し，計算機実験から，既存の単独の手法を適用する場合と比較して，正答率や F 値において改善が見られることを示します．また，出力されたモデルを解釈することで，実際のビジネスに適用可能な方策について検討しています．

以下では，紙幅の都合上，詳細については割愛しますが，問題定義やモデルの概略を説明するとともに，計算結果の一部を紹介します．

もりた　ひろゆき
大阪府立大学 現代システム科学域
〒 599–8531 大阪府堺市中区学園町 1–1
morita@eco.osakafu-u.ac.jp

2. 分類問題の定義と提案手法

データマイニングをビジネスに応用する場合，最も難しい点の一つは適切なデータセットを作成することにあります．応用する分野によっては，すでにクラスや説明変数が決まっている場合もありますが，ビジネス応用の場合，生データから必要な経営課題を解決するために，それに適した目的変数を作成し，またそれを説明する変数も整備しなくてはなりません．この点は，単に分析アルゴリズムにだけ精通していたとしても，如何ともしがたい点であるとともに，効果的な利用を行ううえでは必要不可欠な点であるとも言えます．

まずは，利用可能なデータに対して，さまざまな角度から基礎分析を行った結果を踏まえ，最終的には，商品ジャンルを六つに集約して，顧客ごとの購買金額を説明変数として k-means 法で七つの顧客セグメントを識別しました．そして，その七つの顧客セグメントの年度間における顧客の移動に着目し，その中から直近の課題として，以下のような二つの移動に関する分類問題を定義することにしました．

1. 育児セグメントからファッションセグメントへと移動する購買行動
2. 雑貨セグメントを維持する購買行動

基礎分析から，購買額への影響の大きなものがファッションと雑貨の購買であり，これらは両方を購買する顧客セグメントが確認される一方で，独立して購買する傾向がありました．またファッションでは，子供服から購買を開始した顧客の中で，自分の服を購買するようになる顧客群が確認されました．そこで，上記の行動に対して，移動（または維持）するか，しないかという二つのクラスを設定して分類問題として最終的には定義し，有効だと思われる説明変数を整えて，分類アルゴリズムの入力とします．

この研究では，頑強性とモデルの可読性を重視して，三つの既存アルゴリズムを 2 段階で構成したアンサンブルアルゴリズムを提案します．第 1 段階では，ロジスティック回帰モデルと決定木モデルという異なる特

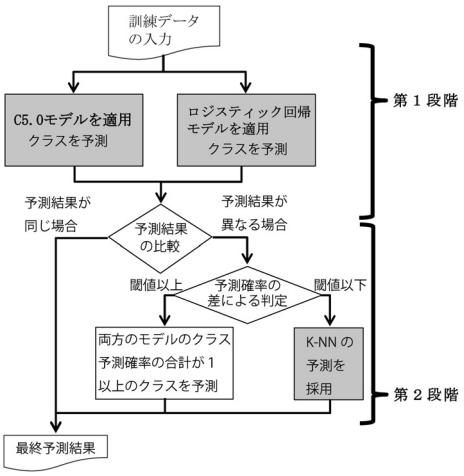

図 1　アルゴリズムの概要

性をもつアルゴリズムで判定を行い，分類予測が異なったレコードに関して，第 2 段階として k-近傍法（以下，k-NN）で再判定させるという方法です．基本的なアイディアとしては，ロジスティック回帰モデルと決定木モデルで，オーバーフィッティングを避けながら異なる角度から予測を行い，異なる予測結果の場合は，k-NN によって補完するということになります．細かな点で言うと，第 1 段階の予測は確率で計算されるわけですが，その確率の大きさや二つのアルゴリズムの確率の相違によって，2 段階目の k-NN の適用を変化させる工夫を研究のオリジナリティとして提案しています．全体の簡単なフローは図 1 のようになります．以下では，その計算結果について紹介します．

3.　実データに対する計算機実験の結果と考察

具体的な計算結果として，前述の育児セグメントからファッションセグメントへと移動する購買行動要因を明確化するため，ある年度に育児セグメントに存在する顧客のうち次年度にファッションセグメントに移動する顧客のクラス（以下，移動クラス）と，利用を休止してしまう顧客のクラス（以下，休止クラス）を選択して，これらを分類する問題のデータセットを作成しました．各アルゴリズムのパラメータについては，予備実験の結果から適切と思われる値に設定し，訓練データ 7 割，テストデータ 3 割になるようにシードを変えて，5 種類のランダムサンプリングしたデータセットを作成しました．各手法を適用した結果の平均値をまとめたものが，表 1 になります．表中の C5.0 は決定木モデルを，LR はロジスティック回帰モデルのことを意味しています．また適合率，再現率，そして F 値

表 1　評価値の計算結果

	C5.0 単独	LR 単独	提案手法
正答率	60.53%	61.14%	61.28%
適合率	60.77%	62.55%	61.60%
再現率	60.74%	56.66%	61.12%
F 値	60.73%	59.45%	61.36%

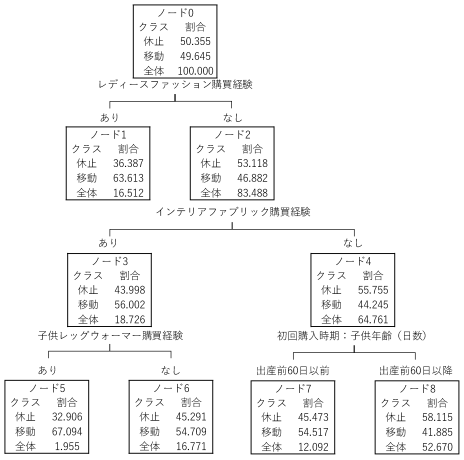

図 2　決定木の上位の分岐

については，移動クラスの値を計算しています．表から，正答率と F 値の両方において，若干ではありますが提案手法の結果がよいことがわかります．C5.0 は適合率と再現率がバランスしていますが，LR は若干適合率を高めるような予測をしています．しかし，提案手法ではこれらをバランスさせて結果を安定させているように見えます．図 2 は，決定木モデルから得られた上位の分岐部分を図示化したものです．図中のノード内の数値は，移動と休止についてはノード内のそれぞれのクラスの割合を，また全体については，全体数に対するノードに該当している顧客数の割合を百分率で表しています．これらを解釈することによって，具体的なプロモーション施策を検討することができます．たとえば，当然ながらレディースファッションの購買経験は，強く影響していることがわかりますし，それらの購買経験がなくても，出産前 60 日より前に，子供の服を用意する行動を起こすことは，ファッションに対する意識の高さを表している徴候かもしれません．

参考文献

[1] トレバー・ヘイスティ，ロバート・ティブシラニ，ジェローム・フリードマン（杉山将ほか監訳），『統計的学習の基礎―データマイニング・推論・予測―』，共立出版，pp.400–402, p. 536, 2014.
[2] 荒木雅弘，『フリーソフトではじめる機械学習入門』，森北出版，2014.