オンラインショッピングサイトにおけるデータマイニングの適用

川崎　明彦† 佐藤　嘉則† 森田　豊久†
(株)日立製作所システム開発研究所†

1. はじめに

本稿では、オンラインショップへのデータマイニング技術適用事例について述べる。

近年、インターネットは企業と顧客とを結ぶ新たな顧客チャネルとして広く認知されてきている。ネット上のオンラインショップでは、ショップ内での行動など従来の顧客チャネルでは取得困難であった情報を比較的容易に得ることができる。しかし、それら顧客に関する情報の全てを人間が理解するのは困難である。そのため、データマイニングのようなデータ理解をサポートする技術が重要になってきている。

本稿では、オンラインショップにおける顧客分析に、if-then 形式のルールを用いたデータマイニング適用手法を示す。実データを用いた実験の結果、リピータおよび優良顧客を上手く説明できるルールを得た。

2. ルールを用いた特徴抽出

図 1 のような if-then 形式のルールを用いて、データ集合の特徴を表現する。if 部を条件部と呼び、1 個以上の「変数=カテゴリ値」の組み合わせであるとする。また、個々の「変数=カテゴリ値」を条件節と呼ぶ。then 部を結論部と呼び、1 個の「変数＝カテゴリ値」であるとする。

結論部によって定義されるデータ集合に対し、それを上手く説明できる条件部を探索することで、結論部の特徴を表すルールを生成する。

```
if   来店頻度 = 大  &  平均購入額 = 中
then  顧客ランク＝ 優良顧客
```

図 1．if-then ルールの例

2.1. 数値変数のカテゴリ化

カテゴリ値とは、「正常」「異常」等のような名義値を意味する。数値変数は、レンジを任意個数の区間に分割し、変数値を「小」「中」「大」などのカテゴリ値に変換して扱う。区間分割は、値域等分割、事例数等分割などデータの特性や特徴抽出目的に応じて決定する。

An Application of Data Mining on Online Shopping Site
Akihiko Kawasaki† , Yoshinori Sato† , Toyohisa Morita†
Systems Development Laboratory, Hitachi, Ltd. †

2.2. ルールの評価尺度と生成方法

if X then Y というルールにおいて、条件部 X が結論部 Y を上手く説明できているか次式で評価する。評価尺度 μ の値が大きいルールほど結論部を上手く説明できているとする。

$$m = P(X)^b \cdot P(Y \mid X) \log\left\{\frac{P(Y \mid X)}{P(Y)}\right\}$$

P(X)は条件部 X の出現確率、P(Y)は結論部 Y の出現確率である。P(Y|X)は結論部 Y の条件部 X に対する条件付確率である。

ここで、P(X)は条件部の一般性を表しているためカバー率と呼ぶ。P(Y|X)は条件部から結論部が出る確からしさを表しているためヒット率と呼ぶ。β はカバー率を調整するパラメータであり、0 以上 1 以下の値とする。

結論部を上手く説明できる条件部を探索するため、指定された条件節数までの「変数=カテゴリ値」の組み合わせ全てについてルールを評価し、評価尺度の値が大きなルールを抽出する[1]。

3. オンラインショップへの適用

上記手法をオフィス家具販売のオンラインショップに適用し、顧客の特徴抽出を行った。ルールを用いることで、可読性が高くデータ理解に適した特徴抽出を目指した。対象となる顧客はリピータおよび優良顧客であり、それぞれ購入回数、購入金額で定義した。

3.1. 分析に使用するデータ

分析に使用するデータは顧客属性情報、商品属性情報、商品販売履歴、Web アクセスログである。

この中で、Web アクセスログはインターネットを介した顧客チャネルに特有のデータである。サイト内で閲覧したページ、ページの閲覧に要した時間、エラー情報などを取得することで、オンラインショップ内における顧客の行動を追跡できる。そのため、従来の顧客チャネルと異なり、商品購入以外の行動についても顧客の特性を表す情報として利用することができる。

3.2. リピータ分析

集計期間中に 2 回以上商品を購入した顧客をリピータ、それ以外の顧客をノンリピータと定義し、前述のルール生成手法を用いて全顧客の中からリピータ特有のルールを抽出した。

ただし、集計締め切り間近に初めて商品を購入した顧客は、集計締め切り後に 2 回目の商品購入をする可能性があり、リピータかノンリピータ判別できないため分析から除外する。

例えば、図 2 の C のような顧客は、締め切り後の期間 T に商品を購入してリピータとなる可能性がある。このような顧客をノンリピータとして扱った場合、本来リピータである顧客がノンリピータとして扱われ、リピータ固有の特徴が抽出されない可能性がある。

そこで典型的なリピータの購入間隔を T 時間とし、1 回目の商品購入から T 時間を過ぎると商品購入の可能性が低くなると仮定する。ここでは、T を求めることでリピータかノンリピータか判別できない顧客を分析から除外する。

典型的なリピータの場合、1 回目の商品購入から T 時間まではノンリピータであると判定出来ないため、集計締め切り時点から T 時間前までの間に初めて商品を購入した顧客を分析から除外する。今回 T にはリピータ全体の 95%が該当する購入間隔を使用する。

また、リピータとノンリピータを比較する際には、両方が同じ条件の変数を使用する必要がある。例えば、リピータはその定義からノンリピータに比べて購入回数が多く、ノンリピータよりも購入金額が高くなる可能性があるため、その変数を比較に使用できない。リピータとノンリピータの違いは、商品購入が継続したか、1 回で終わってしまったかであり、初回の購入についてはリピータもノンリピータも条件は一緒である。そこで、ルール生成で使用する変数には、両者を平等な立場で比較できるように初回購入金額、初回購入商品など初回購入時のものを用意した。

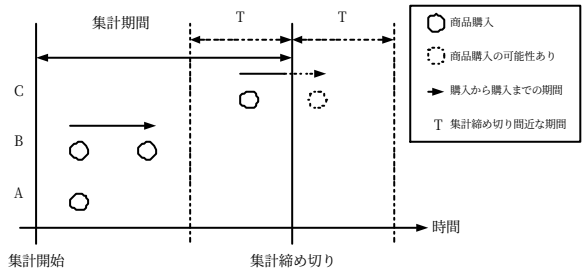

図 2．分析から除外する顧客

ヒット率を重視し、β=0、最大条件節数=3 という条件設定でルール生成を行った結果、リピータの特徴として表 1 の条件部が得られた。

顧客中に含まれるリピータの割合は 38.4%であり、表 1 のヒット率を見れば分かるように、生成されたルールは高い確度でリピータを説明していることが分かる。

評価上位の条件部には初回購入商品に関する条件節が含まれず、初回購入金額や初回購入個数がリピータを上手く説明している。

表 1．リピータ分析ルール生成結果

No	条件部	カバー率	ヒット率
1	住所=東京 & 初回購入金額=大 & 初回購入個数=小	0.198	0.667
2	初回購入金額=大 & 初回購入個数=小 & 初回購入時間帯=9-17 時	0.242	0.636
3	初回購入日=火曜 & 平均商品閲覧数=小	0.319	0.552
:	:	:	:

3.3. 優良顧客分析

集計期間中に一定金額以上商品購入をした顧客を優良顧客と定義し、全顧客の中から優良顧客特有のルールを抽出する。

リピータ分析同様β=0、最大条件節数=3 という条件設定でルール生成を行った結果、優良顧客の特徴として表 2 の条件部が得られた。顧客中に含まれる優良顧客の割合は 20.2%であり、生成されたルールは高い確度でリピータを説明していることが分かる。

条件節には住所や商品カテゴリに加え、訪問傾向などオンラインショップ上での行動に関する変数がリピータの特徴として含まれており、行動に関する情報が重要であることがわかる

表 2．優良顧客分析ルール生成結果

No	条件部	カバー率	ヒット率
1	住所=東京 & 訪問日=月曜 & 訪問時間帯=9-17 時	0.113	0.500
2	訪問日=月曜 & 購入商品カテゴリ=A	0.101	0.471
3	訪問日=月曜 & 訪問時間帯=9-17 & 訪問間隔=短	0.115	0.462
:	:	:	:

4. まとめ

ルールを用いた特徴抽出手法をオンラインショップの顧客分析に適用し、リピータおよび優良顧客の特徴抽出を行った。データの用意および変数選択を厳密に行った上でルール生成を行い、リピータおよび優良顧客を上手く説明できるルールを得た。

参考文献

[1] 芦田、前田、高橋：データマイニングにおける特徴的ルール生成方式，情報処理学会第 50 回全国大会，1995

[2] 牧、前田、内田、中島：ルール生成に基づくデータマイニングの LSI 不良解析への適用、情報処理学会第 52 回全国大会，1996