1B-4

naive bayes を用いたネットユーザのセグメンテーションに関する一考察

坂巻　英一†
公立大学法人宮城大学†
鈴木邦成††
文化ファッション大学院大学††

1．はじめに

如何にして効率的に顧客を店舗に誘導し，顧客満足度を高めリピート顧客へと育て上げるか，がネットショップ運営会社各社にとって重要な経営課題となっている．通常，ネットショップにおいて顧客をショップへ誘導するためには，バナー広告に代表されるディスプレイ型広告が利用されることが多い．ところが，出稿先のサイトによっては，アクセス数は多いものの，購買に結びつく割合が小さい，等広告効果にばらつきがある．そのため，広告を出稿する側にとっては，たとえアクセス数が少なくてもコンバージョン率の高いサイトへ出稿したい，と考えるのが自然であろう．ネット広告の出稿においても TVCM 等と同様，費用対効果が重要なのである．本研究では，単純ベイズ分類器(naive baysian classifier)を利用して顧客を収益性の高いグループとそうではないグループに分類した上で，収益性の高い顧客が流入してくるサイトがどのサイトであるかを，単純ベイズ分類機を用いて把握する方法を提案することを試みる．

2．研究の背景

インターネットの普及とともに様々な商品がネットショッピングで入手できるようになり，ネットショッピング市場は年々拡大する傾向にある．こうした中，ネットショップの運営会社各社は利用者増加を促進するために多様な広告を展開している．このうちバナー広告を始めとしたプル型広告は広く利用されている[1]．一方で，サイトの中には，ネットショップ運営会社に利益をもたらしてくれる優良顧客を多数誘導してくれるサイトもあれば，そうではないサイトが存在するのもまた事実である．このうち，多くのユーザを誘導できる大手ポータルサイト等は広告出稿料も割高であることが多く，広告を出稿する側からすれば，出稿量に見合った収益を得ることができるサイトを見つけ出した上で広告出稿をすることが必要になる．

インターネット上における広告効果に関する研究はインターネットがビジネスで活用され始めた 1990 年代後半以降，盛んに行われてきた．松田[2]らはバナー広告への単純接触がその後の商品評価や購買意欲へ与える影響について，商品カテゴリー，バナー広告の反復露出回数，提示位置等の要因が消費者の商品評価，購買意欲にどのような影響を与えるかを明らかにしている．また，太駄[3]はバナー広告に対するクリック率は一般的に低い傾向にあり，バナー広告の効果が限定的なものである，との報告を行っている．

店舗に多くの収益をもたらしてくれるネットショップにとっての優良顧客がどこのサイトから流入してくるのか，を把握する手法のひとつに単純ベイズ分類器(naïve Baysian classsifier)の利用が考えられる．単純ベイズ分類器とは教師あり学習の設定のもとでベイズの定理を用いた単純な確率的分類器である．本研究では，単純ベイズ分類器を利用して優良顧客の流入元を把握する仕組みを構築する方法を提案する．

3．本研究における提案手法

本節では，単純ベイズ分類器を用いて，ネットショップにおける優良顧客の流入元を把握する仕組みの構築方法について説明する．

3.1 単純ベイズ分類器とは

単純ベイズ分類器とは，事象間における極めて強い独立性を仮定した上で，ベイズの定理を用いて事象を特徴の類似したもの同士に分類することを目的とした確率的分類手法である．

単純ベイズ分類器に関する初期の頃の研究としては，Domingos ら[4]が挙げられる．Domingos らは単純ベイズ分類器を用いた分類に関する効率性に理論的な理由を示している．また，McCallum ら[5]は単純ベイズ分類器において広く利用されている多項モデルとベルヌーイモデルについてモデルの予測精度に関する比較を行っている．以下に顧客セグメントを教師情報として使用し，単純ベイズ分類器に学習させることで，優良顧客が多く流入してくる WEB サイトを把握する仕組みを構築する方法を提案する．

3.2 モデル構築法

今，

- i　WEB サイト番号(i=1,2,…,I)
- I　顧客の流入元となる全ての WEB サイト数
- m　顧客セグメント(m=1,2,…,M)
- M　分析対象となる全顧客数

WEB_i　i番目のWEBサイト

$SEGMENT_m$　m番目の顧客セグメント

とした時, WEBサイトから流入してくる顧客がm番目のセグメントに属する確率を

$$p(SEGMENT_m \mid WEB_1, WEB_2, ..., WEB_i, ..., WEB_I) \quad (1)$$

で表すことにする. 今, (1)式に対してベイズの定理を適用すると, (1)式は

$$p(SEGMENT_m \mid WEB_1, WEB_2, ..., WEB_i, ..., WEB_I)$$
$$= \frac{p(SEGMENT_m)\, p(WEB_1, WEB_2, ..., WEB_i, ..., WEB_I \mid SEGMENT_m)}{p(WEB_1, WEB_2, ..., WEB_i, ..., WEB_I)}$$
$$\propto p(SEGMENT_m)\, p(WEB_1, WEB_2, ..., WEB_i, ..., WEB_I \mid SEGMENT_m) \quad (2)$$

と書くことができる. 今, 流入元となるWEBサイト間に完全な独立性があると仮定すると

$$p(WEB_1, WEB_2, ..., WEB_i, ..., WEB_I \mid SEGMENT_m) = \prod_{i=1}^{I} p(WEB_i \mid SEGMENT_m) \quad (3)$$

のように変形できる.

(3)式を(2)式へ代入すると,

$$p(SEGMENT_m \mid WEB_1, WEB_2, ..., WEB_i, ..., WEB_I) \propto p(SEGMENT_m) \prod_{i=1}^{I} p(WEB_i \mid SEGMENT_m) \quad (4)$$

となる. ここで, $p(WEB_i \mid SEGMENT_m)$を定式化する. 今, $T(WEB_i, SEGMENT_m)$を$SEGMENT_m$に属する顧客がWEBサイトiから流入した回数とする. この時, $p(WEB_i \mid SEGMENT_m)$は

$$p(WEB_i \mid SEGMENT_m) = \frac{T(WEB_i, SEGMENT_m)}{\sum_{i=1}^{I} T(WEB_i, SEGMENT_m)} \quad (5)$$

によって定式化することにする.

ところで, 全ての顧客セグメントにおいて, 分析対象となるWEBサイトからの顧客流入があれば(5)式を計算することは可能であるが, 顧客セグメントの中には分析対象となる全てのWEBサイトから顧客の流入がない場合もある. こうした場合, (5)式の分子はゼロになってしまい(5)式で定式化される確率を計算することができなくなる. こうした問題はゼロ頻度問題と呼ばれており, 各WEBサイトから流入してくる顧客数に対するスムージングを行うことで, 影響を緩和することが可能になる. 一般的に利用されているスムージングの方法として, 全てのWEBサイトについて流入回数に1を加えるラプラススムージング（Laplace Smoothing）と呼ばれる方法がある. ラプラススムージングを利用すると(5)式は(6)式のように書き換えることができる.

$$p(WEB_i \mid SEGMENT_m) = \frac{T(WEB_i, SEGMENT_m) + 1}{\sum_{i=1}^{I} \left(T(WEB_i, SEGMENT_m) + 1\right)} \quad (6)$$

ラプラススムージングを利用すると, $p(WEB_i \mid SEGMENT_m)$がゼロにならないようにすることが可能になる. そこで, 本研究では$p(WEB_i \mid SEGMENT_m)$の計算に(6)式を利用することにする.

更に, ネットショップに来店する顧客は通常様々な流入元から来店する傾向にあるので, $\prod_{i=1}^{I} p(WEB_i \mid SEGMENT_m$は非常に小さな値になることが多い. そのため, (4)式を計算すると, 計算機がアンダーフローを起こす可能性があり, 計算上のアンダーフローを回避するために, (4)式の両辺に対して自然対数をとることにする. (4)式の両辺に自然対数をとると(4)式は(7)式のように書くことができる.

$$\log p(SEGMENT_m \mid WEB_1, WEB_2, ..., WEB_i, ..., WEB_I) \propto \log p(SEGMENT_m) + \sum_{i=1}^{I} \log p(WEB_i \mid SEGMENT_m) \quad (7)$$

この結果, WEBサイトiから流入してきた顧客が属するセグメントは, 関数$f(x)$が最大になるようなxをargmax(x)と書くことにすると,

$$\arg\max\left(\log p(SEGMENT_m \mid WEB_1, WEB_2, ..., WEB_i, ..., WEB_I)\right) = \arg\max\left(\log p(SEGMENT_m) + \sum_{i=1}^{I} \log p(WEB_i \mid SEGMENT_m)\right) \quad (8)$$

によって与えられることになる. 本研究では, (8)式を満たす顧客セグメントmをWEBサイトiから流入してくる顧客が属するセグメントとみなし分析を行うことにする.

参考文献

(1) 新井 亨, メディアプランニングにおけるインターネット広告の役割, 愛知学院大学論叢商学研究, 46(1/2), p.67-81 (2005)

(2) 松田憲・平岡斉士・杉森絵里子・楠見孝, バナー広告への単純接触が商品評価と購買意図に及ぼす効果. 認知科学, 14(1), p.133-154（2007）

(3) 太駄健司,「図解インターネット広告―実務にかかせない基本的な知識から, 効果測定の最新情報まで―」, 翔泳社(2005)

(4) Domingos, P. and Michael, P., "On the optimality of the simple Bayesian classifier under zero-one loss" , Machine Learning , p.103-137, Kluwer Academic Publishers Hingham (1997)

(5) McCallum, A. and Nigam, K., A Comparison of Event Models for Naive Bayes Text Classification, AAAI-98 Workshop on Learning for Text Categorization(1998)

How to classify customers on internet shop with use of naïve baysian classifier

†Yoshikazu, SAKAMAKI　‡Miyagi University

††Kuninori, SUZUKI　Bunka Fashion Graduate University