Webアクセスログを用いた革新性推定手法の推定精度向上に関する一考察

中村美穂†　岸本康成†　市川裕介†　佐藤宏之†　小林透†
日本電信電話株式会社　ＮＴＴ情報流通プラットフォーム研究所†

1. はじめに

Web ユーザの増加に伴い，Web 上の行動履歴データの量が増大し，結果として行動履歴データのマーケティング活動への利用が期待されている．一方，ユーザの価値観が多様化した結果，マーケティング活動ではユーザの心理的な属性（例えば価値観）が重要視されており，今後行動履歴データを用いて Web ユーザの心理的な属性を推定する技術がマーケティング活動において重要になると予想される．

我々は，イノベータ理論[1]で言及されているユーザの革新性に着目し，行動履歴データから革新性を推定する手法の確立に取り組んでいる．革新性とは，ユーザが新しく世の中に出たアイテムに対する採用の積極性を表すユーザの特性である．これまでの研究では，提案手法を行動履歴データに適用した結果，グループの推定結果がひとつのグループ（Majority)に偏るという問題を明らかにした[2]．本論文では，当該問題に対する対策案と検証結果について報告する．

2. 革新性推定手法の概要

実世界のユーザを対象にした調査結果[1]を参考に，下記の仮説を提案する．さらに，仮説は(1)式のように定式化することができる．

仮説：革新性に応じて，各 Web ページへのアクセス回数が異なる．

$$V_i \cdot F(V_i|I_i) = I_i \quad \cdots(1)$$

V_i: ユーザ i の各 Web ページへのアクセス回数を要素とする行動ベクトル．

I_i: ユーザ i の革新性．本研究ではイノベータ理論[1]を参考にユーザを以下の 4 グループで定義する．{Innovator(I), Early Adopter(EA), Majority(M), Laggard(L)} Innovator が最も革新的で，Laggard が最も保守的なグループに該当する．

$F(V_i|I_i)$：行動ベクトル V_i からユーザ i の革新性 I_i を推定するモデル．行動ベクトルと革新性が関連づいた一部のユーザのデータ（学習データ）を用いて推定する．

モデル $F(V_i|I_i)$ を構築することができれば，革新性が未知のユーザ j の革新性 I_j を，V_j から推定することが可能となる．

本研究では，推定モデル $F(V_i|I_i)$ の構築は V_i が Web ページの数に伴い高次元になる可能性が高いことを考慮し，高次元の線形識別を得意とする Support Vector Machine(SVM)[3]を用いた．

3. Sampling 手法の提案

1 章で述べた問題の原因の一つとして，モデルの推定に用いた学習データの人数の偏り imbalanced data sets problem(IDS 問題)[4]が考えられる．IDS 問題への対策として，学習データの各グループのデータ数の偏りを調整する方法(sampling 手法)が提案されている[4]．本研究では，ランダムな sampling 手法が各種データに適用した結果，平均的に良い結果を残すという Hulse らの報告[5]を参考に，3 グループ以上に適用可能な手法 Multi-class Random Over/Under Sampling(MROS/MRUS)を提案し，手順を示す．

(1)sampling 後のデータ数の比率 r を定める（r は 0～1 の実数）．MROS の場合，比率 r はデータ数が最も多いグループ(図 1 の A)のデータ数に対する最も少ないグループ(図 1 の C・D）のデータ数に該当する．

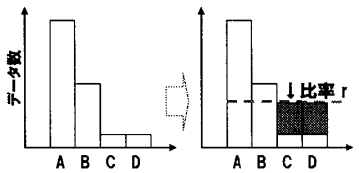


図1. MROSの概要

(2)実際のデータ数と比率 r から sampling 後のデータ数を導出し，データ数が不足（過剰）しているグループに対して，元来含まれていたデータの中からランダムに抽出したデータを複製（削除）して，データ数の比率が(1)で定めた比率が r になるよう調整する．

4. 検証

提案手法を行動履歴データを用いて検証した．

4.1 実験データ

ビデオリサーチインタラクティブ社が収集した Web 視聴データと特性調査データのうち，電気製品を扱う某 EC サイトに 2007 年 1 月から 8 月にアクセスしたユーザ（225 名）を対象にした．Web 視聴データは，ユーザが閲覧した Web ページの URL を記録したデータで，特性調査データは Web 視聴データの収集ユーザに対して 2006 年 11 月にアンケートを実施し取得したデータである．

A Study of Profiling Method of User's Innovativeness with Access Log
† Miho NAKAMURA, Yasunari KISHIMOTO, Yusuke ICHIKAWA, Hiroyuki SATO and Toru KOBAYASHI : NTT Information Sharing Platform Laboratories, NTT Corporation

4.2 実験方法

(1)特性調査データを用いてユーザの革新性を導出した．なお，革新性の導出ロジックは，イノベータ理論[1]の各グループの定義を参考にして定めた．導出したデータの各グループのユーザ数は，下記の通りであった．

I：29人，EA：49人，M：113人，L：34人

(2)各 Web ページへのアクセス回数を集計し，行動ベクトルを算出した．なお，本検証では下記の条件に該当するデータを分析対象外とした．

・1名以下のユーザしかアクセスしなかった Web ページへのアクセスデータ

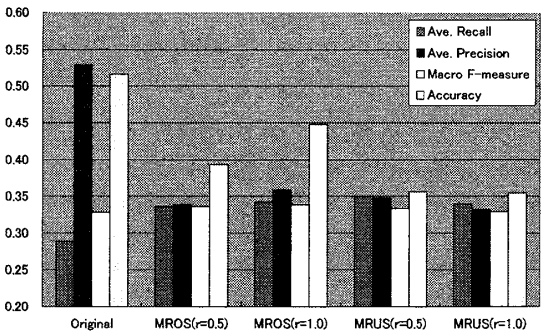
・分析対象サイトへの総アクセス回数が，3回以下のユーザのアクセスデータ

(3)(1)で導出したデータと，(2)で生成したデータに SVM を適用し，One-Leave-Out Cross Validation(CV)法で推定モデルを構築した．なお，推定モデルを構築する際には，当該データに対し3章で述べた MROS と MRUS を適用し，データ数の調整を行った．sampling 比率 r は，0.5 と 1.0 に設定した．また，本検証では，sampling による推定モデルの揺らぎを抑える為，各 r に対して One-Leave-Out CV を5回行い，各試行の平均値を結果とした．SVM の実装には LIBSVM[6]を用いた．カーネルは予備実験で最も良い値を示した RBF(Radial Basis Function)を用い，各種パラメータは最も Macro F-measure[7]が高くなるようグリッドサーチにより探索した．

本実験では，各グループの平均値に該当する Averaging-Precision ， Averaging-Recall ， Macro F-measure と，各手法の推定精度に該当する Accuracy を算出した．

5. 結果・考察

sampling 処理後の各推定精度を図 1 に示す．比較の為，sampling 処理を行わなかった場合の推定精度を original として図1に併せて示す．



※Original の Macro F-measure は4グループのうち1グループの値を算出できなかったので，3 グループの平均値を算出した．

図 1. Sampling 後の推定精度

図1より，sampling を行ったほうが Recall と Macro-F. は高い値を示すことを確認した．これは sampling を行うことにより，どのグループのユーザも満遍なく推定できるようになった事を意味する．以上より，IDS 問題に対して提案手法が有効である可能性が高い事を確認した．

更に，各グループでの sampling の効果について考察する為，図1の MROS(r=1.0)に関して各グループの Recall とその変化量を表1に示す．結果，提案手法の適用により Majority(M)以外のグループの Recall が約 10%から 30%上昇していることを確認した．M の Recall が減ったのは，正しく M と推定されたユーザの数が減った為であり，M の中には他グループと類似した行動をするユーザが含まれていた事を示している．

表 1. 各グループの Recall の変化

	Original	MROS(r=1.0)	変化量[%]
I	0.0	29.0	+29.0
EA	2.0	11.4	+9.4
M	96.5	68.3	-28.2
L	14.7	28.2	+13.5

Recall の上昇は，該当するユーザをもれなく抽出できるという点で重要である．例えば，Recall が高い推定技術は広告配信先のユーザを抽出する技術として活用可能である．なぜなら，広告配信は，1通あたりの送信コストが比較的安価のため，配信先のユーザに少々ノイズとなるユーザが含まれていても反応する可能性の高いユーザをもれなく抽出し配信できる方が望ましいからである．今後，本推定手法はより各グループの Recall を高めることにより，広告配信サービスへの適用が期待される．

6. まとめ

推定結果がひとつのグループに偏る IDS 問題に対して，マルチクラスの学習データに利用可能な sampling 手法を提案した．さらに，行動履歴データを用いて提案手法の有効性を検証し，提案手法が IDS 問題に対して有効である可能性が高いことを確認した．今後は，行動ベクトルの生成方法などを検討し，推定精度の向上に取り組む予定である．

参考文献

[1] E. M. Rogers: Diffusion of Innovations, 5th edition, The Free Press: New York(1982)(三藤 利雄訳:イノベーションの普及，翔泳社(2007))
[2]中村美穂ら: アクセスログを用いたユーザの革新性推定方式に関する一考察,SIG-KBS, pp.27-32(2008)
[3] Vapnik, V: Statistical Learning Theory, Wiley, New York(1998)
[4] S.Visa et al.: Issues in Mining Imbalanced Data Sets-A Review Paper, in Proc. of the 16th Midwest Artificial Intelligence and Cognitive Science Conf., pp.67-73(2005)
[5] J.V.Hulse et al.: Experimental perspectives on learning from imbalanced data, in Proc. of the 24th Int. Conf. on Machine learning, pp.935-942(2007)
[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[7] Y. Yang: An Evaluation of Statistical Approaches to Text Categorization, Information Retrieval, vol.1, 1-2, pp.69-90(1999)