オーラルセッション ― フルペーパー

# インターネットテレビにおけるユーザの視聴行動分析

## ― 継続・離脱分析とユーザアクティブ度の定量化 ―

株式会社サイバーエージェント
和田 計也

株式会社サイバーエージェント
福田 一郎

**要約**

本研究の目的は，弊社がインターネット上で展開している一テレビサービスを例にとり，サービスの継続・離脱分析を翌日単位ではなくて翌週単位で行う方法及び，単なる継続・離脱分析に留まることなく同時にユーザのアクティブ度を定量化することもできる方法を提案するものである。具体的には分析モデルとしてmultivariate adaptive regression splines(MARS)を用い，翌週アクティブ化日数を目的変数とした二項分布を仮定することで翌週単位での継続・離脱分析を実現しており，二項分布での成功確率pをユーザのアクティブ度と定義することで定量化を実現している。また実ビジネス上のデータでよく直面する，正規分布とは懸け離れた分布のデータであってもMARSを用いることでうまくモデル化することが可能である。

**キーワード**

Web，テレビ，multivariate adaptive regression splines(MARS)，離脱分析

## I. はじめに

インターネットの発展に伴い，インターネット上で展開されるサービスは従来の単純なWebサイトからFacebookやtwitter等のSNSに至るまでたくさんの種類が出現してきた。近年ではYouTubeに代表されるような動画サービスも80%以上の人が利用したことがあるほど一般的になってきている(MMD研究所 2015)。最近ではNetflixやHuluなど，インターネット上でテレビ番組を視聴できるようなサービスも展開されるようにもなってきた。2016年4月より弊社でもインターネットテレビサービスを手がけているが，このようなインターネット上で展開されているサービスは据置型のゲームやテレビなどと比べて初期費用がかからないなどの理由により導入が簡単な反面，サービス利用継続率が時間と共に低下していくという問題を抱えている。(野島，2014)　そのため，ユーザの行動を分析することで離脱しないような要因を発見してサービス改善に役立てることができれば，ユーザにとってもサービス運営側にとってもお互い良いことだと言えよう。本研究ではインターネットテレビにおけるユーザの継続・離脱行動を分析すると共に，ユーザのサービスへのアクティブ度を定量化することを目的とした。

本研究で必要とされる要件は以下の3件であった。i) 人が見て解釈可能であること。ii) 正規分布を仮定しないノンパラメトリックな手法であること。iii) 変数選択ができること。これらの要件を満たす最良の手法と思われた，multivariate adaptive regression splines(MARS)(Friedman，1991)を本研究では用いた。Webサービス等で継続・離脱分析を行う場合に翌日の継続離脱を目的変数とした分析がなされることあるが，必ずしも適切な目的変数設定であるとはいえない。翌日にサービスへのログインが無い場合であってもユーザが離脱したとは言い切れないからだ。そこで本研究では翌週のアクティブ日数(≒ログイン日数)を目的変数と設定した継続・離脱分析を行った。また，継続・離脱分析と同時に個々のユーザのアクティビティ度を定量化することも行った。

## II. 先行研究

片岡らはPOSデータにClassification by aggregating emerging pattern(CAEP)を適用させて優良顧客の離脱予測モデルを構築した(2011)。また, 佐藤らはオンラインソーシャルゲームの行動ログを混合正規分布によるクラスタリングにより, ユーザの離脱傾向を研究した。藤井らはCD購買POSデータを決定木により継続購入するユーザのモデリングを行った。このように, 様々なデータに対して分類器を使いユーザの離脱分析を行う事例は従来から行われている。ChouらはMARSモデルとニューラルネットワークを組み合わせて主に所見(腫瘍サイズ等)から作成された変数を用いて乳がんの術後再発症パターンのマイニングを行った(2004)。ChouらはMARSで変数選択を行い, ニューラルネットワークで予測モデルを構築する組み合わせを試している。このように分類器としてMARSモデルを適応させる事例も行われてきている。

## III. 方法

### 1. データ収集の方法

弊社はインターネット上で, オリジナルの生放送コンテンツや, ニュース, 音楽, スポーツ, アニメなど多彩な番組が楽しめる約20チャンネルをすべて無料で視聴することができるインターネットテレビサービスを2016年4月より提供している。図1に示すとおり, サービスの質向上やユーザの満足度向上等のためにユーザ行動のログをHadoop上のHiveテーブルに蓄積しており, サービスの分析に生かしている。

### 2. 利用したデータ

本研究に利用したデータは表1に示す通り, 弊社で運営しているインターネットTVサービスでの2016年5月1日から2016年6月2日までの期間で約110万人のユーザをランダムサンプリングしたものである。Hadoop上のHiveテーブルに蓄積された行動ログから, データ分析として用いることができるようにHiveQLで集計・抽出・整形を行いデータを得た。

**表―1 データの概要**

| | |
|---|---|
| サービス名 | インターネットテレビサービス |
| 期間 | 2016年5月1日~2016年6月2日 |
| 分析利用ユーザ数 | 約110万人をランダムサンプリング |

### 3. モデル

アクティブ度をp(0.0 ～ 1.0)とすると, 翌週アクティブ日数(0 ～ 7)Yは以下のような二項分布に従うと仮定する。

$$Y \sim Binom(7, p)$$

アクティブ度pは関数f(x)で得られる実数値をlogit変換することで0.0から1.0の範囲の確率値となる。

$$logit(p) = f(x)$$

関数f(x)は線形和の場合だと一般化線形モデルとなるが, 本研究で用いたMARSモデルでは各説明変数の効果量$\beta m$に対し更にhinge関数である$hm(x)$が乗算されたものとなっている。Σの中で計算される変数は表2に示すとおりである。

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \, h_m(x)$$

$$h(x) = \begin{cases} max(0, x - const), & if\ x > const, \\ max(0, const - x), & otherwise. \end{cases}$$

下記Generalized Cross Validation(GCV)の値が最小になるように変数選択が行われる。

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}(y_i - f_\lambda(\chi_i))^2}{(1 - M(\lambda)/N)^2}$$

MARSモデルのfitting はRのearthライブラリ[1)]を用いて行った。

本モデルはある日のアクティブ度が翌週のアクティブ日数を決めるというやや強めの仮定を置いているが, このようなモデルとすることでユーザの継続・離脱に繋がる要因を

示唆できると共に，ユーザ毎にアクティブ度を算出できるようになるのでサービスの現状把握などにも利用することが出来るためビジネス現場で使い勝手の良いモデルとなっている。

### 4. 利用した変数

MARSモデルで翌週のアクティブ化日数をモデリングする際に利用した変数を表-2に示す。翌週アクティブ日数が目的変数であり，それ以外の変数は全て説明変数である。変数の概念図を図2に示す。説明変数は，ある特定の1日の中での行動と，その日からみて前週1週間のアクティブ日数から成る。構築されるモデルはこれらの説明変数から，明日以降の翌週1週間のアクティブ化日数を予測するモデルである。

## IV. 結果

前述の通り，社内のオンプレミス環境にミドルウェアであるHadoopを使って構築しているデータ収集基盤からHiveQLにより実験データを取得した。データの記述統計は表3および表4に示されている。

総視聴時間に着目すると，平均値と中央値とに大きな乖離が見られることと標準偏差が平均値の2倍弱であることから，釣鐘状の正規分布ではない分布形状であることが容易に想像できる。総視聴時間の分布を図3に示すとおり，べき分布と思われる分布形状をしており(ここでは，べき分布であるかどうかは本質的な問題ではないためべき分布性の確認は行っていない。)，この説明変数を用いたパラメトリックな線形モデルを使うべきではない。これが，本研

**表―2 利用した変数一覧**

| 変数名 | 詳細 |
|---|---|
| 翌週アクティブ日数 | アクティブ日数とは30秒以上何らかの番組を視聴した行為があった日数で0~7の値 |
| 前週アクティブ*日数 | アクティブ日数とは30秒以上何らかの番組を視聴した行為があった日数で0~7の値 |
| 総視聴時間 | 1日の視聴時間合計(分単位) |
| 視聴チャンネル数 | 1日の30秒以上視聴チャンネル数　1~20の値 |
| 番組予約数(0~) | 0以上の値 |
| プラットフォーム | スマホアプリ/PCweb |
| 視聴チャンネルカテゴリのダミー変数 | ニュース系，ドラマ系，麻雀系，アニメ系，バラエティ系，音楽系，スポーツ系，その他 |

**表―3 連続値データの記述統計**

| | 平均値 | 中央値 | 標準偏差 | 最小値 | 最大値 |
|---|---|---|---|---|---|
| 翌週アクティブ日数 | 2.81 | 2 | 2.36 | 0 | 7 |
| 前週アクティブ日数 | 2.64 | 2 | 2.40 | 0 | 7 |
| 総視聴時間(分) | 43.80 | 12 | 88.19 | 0 | 1437 |
| 視聴チャンネル数 | 2.06 | 1 | 1.63 | 0 | 20 |
| 予約チャンネル数 | 0.34 | 0 | 2.96 | 0 | 508 |

**表―4 カテゴリデータの割合**

| | TRUE | FALSE |
|---|---|---|
| スマホアプリflag | 86.5% | 13.5% |
| ニュース系番組視聴 flag | 23.7% | 76.3% |
| ドラマ系番組視聴 flag | 11.9% | 88.1% |
| 麻雀系番組視聴 flag | 11.2% | 88.8% |
| アニメ系番組視聴 flag | 51.5% | 48.5% |
| バラエティ系番組視聴 flag | 21.9% | 78.1% |
| 音楽系番組視聴 flag | 12.9% | 87.1% |
| スポーツ系番組視聴 flag | 19.9% | 80.1% |
| その他番組視聴 flag | 18.1% | 81.9% |

究で分布形状によらないノンパラメトリックなMARSを選択した理由の一つである。

**表―5 MARSモデルにおける効果量**

| 変数名 | 効果量 |
|---|---|
| (切片) | – 0.164 |
| プラットフォームPC | – 0.406 |
| アニメ系番組視聴 flag | 0.172 |
| h(35 – 総視聴分) | – 0.018 |
| h(総視聴分 – 35) | 0.002 |
| h(1 – 前週アクティブ日数) | – 0.220 |
| h(前週アクティブ日数 – 1) | 0.347 |
| h(2 – 視聴チャンネル数) | – 0.225 |
| h(視聴チャンネル数 – 2) | 0.055 |
| h(2 – 予約数) | – 0.297 |

MARSモデルにfittingした結果を表5に示す。大まかな傾向として,①スマホアプリで視聴②アニメ系チャンネルの視聴③総視聴時間35分以上④前週アクティブ日数1日以上⑤視聴チャンネル数2つ以上⑥予約番組数2つ以上がユーザのアクティブ度を高めるのに効果的(≒つまり翌週アクティブ日数が増加するということ)であることがわかる。例えばPCでスポーツ系チャンネルとバラエティ系チャンネルと音楽系チャンネルとを合計60分視聴したユーザで前週アクティブ日数が2日,予約数が0のユーザの場合下記のような計算によりアクティブ度0.329(32.9%)であると算出される。

$$\eta = -0.164 - 0.406\times1 + 0.172\times0 + 0.002\times(60-35) + 0.347\times(2-1) + 0.055\times(3-2) - 0.297\times(2-0) = -0.712$$

$$p = \frac{e^{-0.712}}{(1+e^{-0.712})} = 0.329$$

## V. 結論

MARSモデルを用いてインターネットテレビサービスにおけるユーザの継続・離脱分析を行いユーザの離脱につながるような行動を示唆することができた。また,同時に個々のユーザのアクティブ度を定量化する方法を提案した。今後は,週末に視聴率が上がるというテレビ特有の現象を加味して曜日効果をモデルに含めるなどしてモデルの信頼度を高めていきたい。

### 注

1) MARSという名称はSalford Systems社が所有しており商用MARSソフトウェアを販売しているため,MARSの実装系はearthという名称になっていることが一般的である。

### 参考文献

MMD研究所 (2015),“無料動画アプリに関する利用実態調査” https://mmdlabo.jp/investigation/detail_1453.html

野島,中村,遠藤,三上 and 近藤 (2014),“アクションポイント制ソーシャルゲームにおける離脱要因の実証実験による検証”日本デジタルゲーム学会 2014年年次大会予稿集

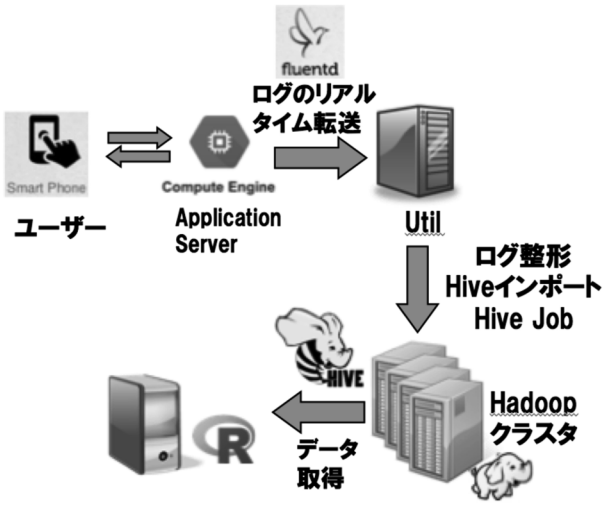
Friedman J. H. (1991),“Multivariate Adaptive Regression Splines”. The Annals of Statistics. 19: 1.

片岡,森田 (2011),“異常検知を利用した優良顧客離脱予測モデル”経営情報学会　全国研究発表大会要旨集 2011f(0), 78-78, 2011

Chou S.M, Lee T.S, Shao Y.E, and Chen I.F (2004),“Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines”. Expert Systems with Applications 27 (2004) 133-142

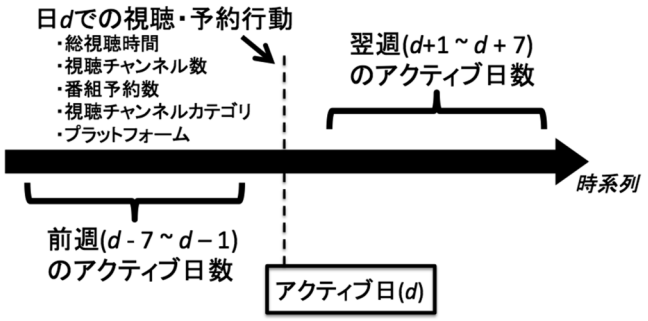
Milborrow S (2011)“earth: Multivariate Adaptive Regression Splines” R packages

**図 1. ログ収集，解析システム概要**



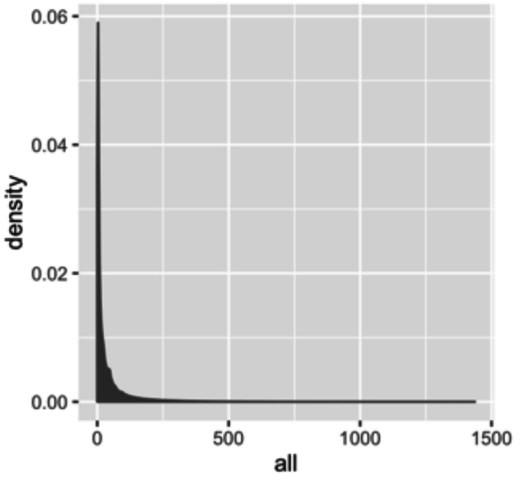

Android, iPhone のネイティブアプリもしくは PC の web ブラウザでの視聴，行動履歴は Google Cloud Platform 上で動いている Application Server 上でログが生成される場合であってもクライアント側で生成される場合であっても一旦 Application Server 上に集約される。そこから Fluentd を介してリアルタイムにログを転送している。転送されたログは Util サーバ上でログ形式のチェックがかかり，形式不備の無いものだけが Hadoop 上の Hive テーブルに一定の間隔でインポートされる。この Hive テーブル内にインポートされたデータを分析サーバに読み込んで分析を行う。

**図 2. 利用変数の概念図**




MARS モデルでの翌週継続・離脱分析に用いた変数を図示した。あるユーザがアクティブ ( 何らかの番組を 30 秒以上視聴した状態 ) になったある一日を d とすると，その日以降 d+1 から d+7 までのアテクィブ化日数が目的変数となる。ある一日 d 中での総視聴時間，視聴チャンネル数，番組予約数，視聴チャンネルカテゴリ，視聴アプリのプラットフォームと，日 d 以前の d-1 から d-7 までのアクティブ化していた日数が説明変数となる。

**図 3. 総視聴時間 ( 分 ) の分布**




対象データの全ユーザで，一日の視聴時間の分布をヒストグラムでプロットした図。X 軸が視聴時間 ( 秒 ) で Y 軸が密度である。正規分布ではなく，ロングテールを形成している冪分布のような形状となっている。