日本オペレーションズ・リサーチ学会
2019年 秋季研究発表会

裏番組を考慮したターゲットごとの視聴率予測

東京工業大学　*山野上勇人　YAMANOUE Yuto
東京工業大学　石田雄基　ISHIDA Yuki
東京工業大学　小茂田岳広　KOMODA Takehiro
05001222　東京工業大学　住谷有規　SUMIYA Yuki
東京工業大学　小泉直人　KOIZUMI Naoto
01405430　東京工業大学　中田和秀　NAKATA Kazuhide

1. はじめに

本研究では，経営科学系研究部会連合協議会主催，平成 30 年度データ解析コンペティションにて株式会社ビデオリサーチから提供された TV × WEB のシングルソースデータ『VR CUBIC』を扱う．これは、ユーザーの TV 視聴ログや CM の放送履歴などが記録されているマルチメディアのデータである．

広告主は自社の商品，サービスに興味を持ちそうな人々（ターゲット）に出来るだけ多く CM を見せることが重要である．そのため，各テレビ番組に対してターゲットの視聴率が予測できれば，広告主は効率的に CM の出稿先を決めることが可能となる．よって，本研究では各番組に対して任意のターゲットごとの視聴率を予測するモデルを提案する．

2. 提案手法

本研究では，ターゲットごとの視聴率の予測を以下のように行う．まず，個人ごとに各番組の視聴確率を算出する．このとき,「標本情報」,「番組情報」,「ライン情報」の 3 つを入力とした機械学習の予測モデルを用いる．次にターゲットに含まれる個人の視聴確率を平均して，その値をターゲットの予測視聴率とする。この際，個人ごとの各番組の視聴確率を精度よく予測するために次の 3 つの工夫を加えた．

2.1. 裏番組の考慮

提供されたデータについて分析したところ，番組の視聴率はその番組の裏番組によって大きく左右することが分かった．しかし，各局の番組放送時刻は一律ではないため，裏番組の組み合わせは複雑で扱いにくい．本研究では，この問題をサンプリングで解決することにし，それをライン抽出法と名付けた．

ライン抽出法の全体イメージを図 1 で示した．番組表を 1 分単位で区切ると，ある時刻のライン上では各局の番組を一意に定めることができる．そのライン上での各局の番組の特徴量を繋ぎ合わせたものを「番組情報」とし，ラインに関する特徴量（時刻，曜日等）を「ライン情報」とする．これらと個人に関する特徴量である「標本情報」を組み合わせたものを入力とし，一つの入力に対してどの局の番組を視聴したか，あるいはどれも視聴しなかったかの多クラスのラベルを付与する．ただし，すべての時刻でのラインを入力に用いると学習データが膨大になり計算に時間がかかるため，ランダムにラインをサンプリングして学習を行う．

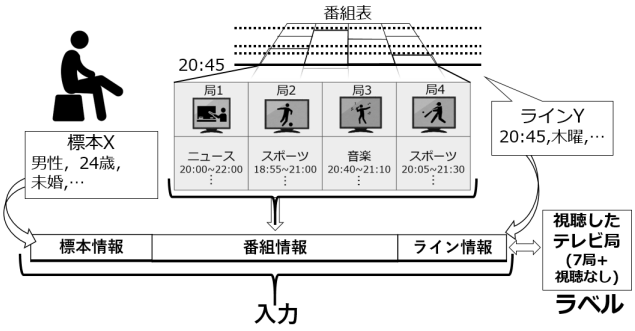

図 1: ライン抽出法の概略図

2.2. 二段階学習法

前述のようにラベル付けを行うとほとんどのデータが視聴しなかったというクラスであり、通常通り学習を行うと視聴したというラベルの学習が不十分になる．その対策として視聴率の予測を二段階に分けて学習する二段階学習法を提案する．一段階目では，個人がテレビを視聴したか否かを判別する二値分類を行う．二段階目では，もしテレ

ビを視聴していたときにどの局の番組を見るかの多クラス分類を行う．よって，個人の局ごとの視聴率はテレビを視聴した確率とその局の視聴確率の積から求まる．なお，この手法を用いるとき各段階ごとに予測モデルを変更することが可能となり，本研究では段階ごとのモデルの組み合わせパターンとして一段階目はランダムフォレスト (RF)[1] と XGBoost(XGB)[2] の内から，二段階目をロジスティック回帰 (LR)[3] とランダムフォレストと XGBoost の内から選択した．

2.3. 新規特徴量の追加

より予測精度を向上させるため，提供されたデータに含まれていない，視聴率に大きく影響を及ぼすと考えられる特徴量を追加した．その一つは各番組の出演者の情報である．この情報を過去のテレビ番組の出演者が網羅されている WEB サイト [4] から抽出した．ただし，次元数を削減するために男性女性それぞれ上位 50 人の人気タレントについてのみ取り扱った．

3. 数値実験

3.1. データ及び実験概要

提供されたデータの TV 接触ログを用いて提案手法の有効性を検証した．学習データは 2017 年 4 月 3 日から 6 月 30 日までの 1 クール分，検証データは同年 7 月 1 日から 14 日の 2 週間分を用いた．また，日ごとのライン数は学習データと検証データでそれぞれ 3,000 本と 1,400 本である．

実験は学習モデルの選定，ある新規番組の各ターゲットごとの視聴率予測の 2 種類である．最初の実験では過去の平均視聴率を予測値とするルールベース手法と提案手法の比較に加え，提案手法に用いる予測モデルの検討を全体視聴率を予測するタスクを通して行う．次に，一つの番組に対するターゲットごとの視聴率を予測させる．なお，今回の実験では相関係数と平均絶対誤差（MAE）を評価指標として用いる．

3.2. 実験結果

1 つ目の実験では予測対象を既存番組と新規番組に分けて検証を行う．ルールベース手法について，既存番組には学習期間の同番組の平均視聴率を，新規番組には学習期間の同時刻同曜日の平均視聴率を予測値として用いる．実験結果を表 1 に示す．実験結果より，一段階目と二段階目どちらも XGBoost を用いた提案手法が最も優れていることが分かる．

2 つ目の実験ではターゲットを予め設定し，ある新規番組についての視聴率予測を行う．実験結果を表 2 に示す．実験結果より提案手法がどのターゲットにおいても精度よく予測できていることが分かる．

表 1: モデルごとの視聴率予測

手法	既存番組		新規番組	
	相関係数	MAE	相関係数	MAE
ルールベース	0.958	0.305	0.798	0.669
RF×LR	0.934	0.392	0.791	0.696
RF×RF	0.962	0.298	0.839	0.621
RF×XGB	0.969	0.270	0.835	0.620
XGB×LR	0.939	0.383	0.787	0.694
XGB×RF	0.968	0.286	0.837	0.619
XGB×XGB	0.971	0.257	0.832	0.618

表 2: 2017 年 7 月，TBS の新規番組「ハロー張りネズミ」(22:00-23:09) のターゲットごとの視聴率予測 (XGB×XGB)

ターゲット	人数	正解	予測	絶対誤差
全体	5,208	5.60%	5.30%	0.30%
男性	2,694	4.01%	3.43%	0.58%
女性	2,514	7.30%	7.30%	0.00%
学生	505	3.01%	3.05%	0.04%
父親	824	4.03%	3.30%	0.72%

参考文献

[1] L. Breiman，“Random Forests，” *Machine Learning*, **45**，pp. 5-32，2001.

[2] T. Chen，C. Guestrin，“XGBoost: A Scalable Tree Boosting System，” *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794，2016.

[3] D. R. Cox，“The Regression Analysis of Binary Sequences，” *Journal of the Royal Statistical Society: Series B (Methodological) banner*，**20**，pp. 215-242，1958.

[4] 価格.com,「価格.com - テレビ紹介情報　［テレビ番組で紹介されたお店や商品の情報］」，https://kakaku.com/tv/，2019 年 6 月 24 日閲覧