7C-02

SNSにおけるバズ予測のための 各利用者が他者の投稿に対してお気に入り登録を行う要因分析

荒澤 孔明† 松川 瞬† 杉尾 信行† 和田 直史† 松﨑 博季†

北海道科学大学 工学部 情報工学科†

1 はじめに

SNSマーケティングにおいて，人々の関心を誘発させる投稿をいかに制作し，いかに発信するかは重要な課題である．特に，ある投稿に対し，人々がどの程度の関心を持つか（バズるか）を予測し，その要因を特定する技術が必要である．

本稿では，SNSユーザごとに，これまでお気に入り登録を行ってきた投稿群の特徴を学習し，これから閲覧する投稿に対して，お気に入り登録を行うか否かを2値分類する予測モデルを，勾配ブースティング決定木に基づき構築する．これにより，ユーザが任意の投稿に対して，お気に入り登録を行う要因の分析が可能になる．

2 お気に入り登録の予測性能に関する実験

2-1 実験環境

協力者はTwitterユーザ8名であり，テスト期間は1ヵ月である．8名の平均テストデータ数は，閲覧後にお気に入り登録を行った投稿が68.4（±72.7）件，閲覧後にお気に入り登録を行わなかった投稿が1717.4（±2393.5）件である．説明変数は表2である．なお，予測には，XGBoost[1], LightGBM[2], CatBoost[3]を用いており，パラメータ調整については，文献[4]を参照されたい．

2-2 実験結果

図1は，学習期間を2〜24ヶ月と変化させた時の予測性能（8人の予測モデルの平均F値）である．XGBoostでは，学習期間が長いモデルほどF値が向上し，CatBoostではその逆の傾向が窺えた．また，LightGBMでは，F値のピークが捉えづらく，学習期間の依存性が高い事も分かった．

表1は，3つの手法における最高性能である．最良のモデルは，CatBoostで6ヶ月間学習を行ったもので，その時のF値は0.645であった．SNSでは，ユーザが気まぐれにお気に入り登録を行う場合もあり，本タスクの難易度を考慮すると，この性能は，十分実用可能であると考察する．

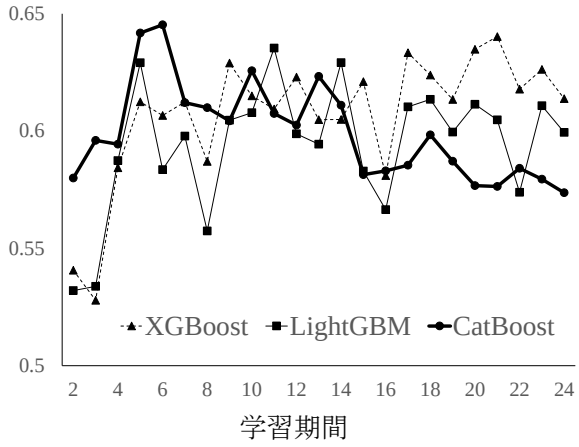

図1 学習期間に依る3手法のF値の変化

3 お気に入り登録の要因に関する実験

3-1 実験環境

ユーザごとに，CatBoostで6ヵ月間学習を行った予測モデルを構築すると，各モデルで，20種類の説明変数の重要度が得られる．これを20次元の特徴ベクトルとし，ユーザ8人に対して，ウォード法によるクラスタ分析を行った．なお，距離の測度には，ユークリッド距離を用いた．

3-1 実験結果

表3の1列目は，8名の予測モデルの20種類の説明変数の平均重要度であり，2・3行目は，2クラスタに分けた際の各々の中心ベクトルである．

第1クラスタでは，あるユーザのある投稿に対するお気に入り登録予測において「その投稿者の過去の投稿にユーザがどの程度お気に入り登録を行ったか（変数 5）」が重要視されている．逆に「その投稿内の単語をユーザが過去にどの程度つぶやいたか（変数 9）」は重要視されていない．ここに該当する4名は「投稿者が誰であるか」が主な要因で，その投稿にお気に入り登録を行う人物らであると解釈できる．

表1 3手法の最高F値（括弧内は学習期間）

	XGBoost (21)	LightGBM (11)	CatBoost (6)
再現率	0.900	0.844	0.817
適合率	0.523	0.529	0.574
F値	0.640	0.635	**0.645**

An Analysis of Factors that Cause Each User to Bookmark Posts in order to Forecast Buzz on Social Network

† Department of Information and Computer Science, Faculty of Engineering, Hokkaido University of Science

表 2 あるユーザが投稿pに対してお気に入り登録を行うか否かを予測するための説明変数

	値域	概説
1	$\{0,1\}$	投稿pに動画像が含まれるか否か
2	$\{0,1\}$	投稿pがユーザ自身への返信か否か
3	≥ 0	投稿pの投稿者のフォロワー数
4	$[0,1]$	過去にユーザが返信した全投稿中，投稿pの投稿者のものが占める割合
5	$[0,1]$	過去にユーザがお気に入り登録した全投稿中，投稿pの投稿者のものが占める割合
6	$[0,1]$	過去にユーザが拡散した全投稿中，投稿pの投稿者のものが占める割合
7	≥ 0	ユーザが過去に発信した投稿内での，投稿pに含まれるタグの出現回数
8	≥ 0	ユーザが過去にお気に入り登録した投稿内での，投稿pに含まれるタグの出現回数
9	≥ 0	ユーザが過去に発信した投稿内での，投稿pに含まれる単語の出現回数
10	≥ 0	ユーザが過去にお気に入り登録した投稿内での，投稿pに含まれる単語の出現回数
11	≥ 0	ユーザが過去に発信した投稿内での，投稿pの投稿者の Bio に含まれる単語の出現回数
12	≥ 0	ユーザが過去にお気に入り登録した投稿内での， 投稿pの投稿者の Bio に含まれる単語の出現回数
13	≥ 0	ユーザが過去に発信した投稿内での，投稿pの投稿者名の出現回数
14	≥ 0	ユーザが過去にお気に入り登録した投稿内での，投稿pの投稿者名の出現回数
15	≥ 0	投稿pに周囲が付与したお気に入り登録数
16	≥ 0	投稿pの投稿者の過去の投稿に，周囲が付与してきた平均お気に入り登録数
17	≥ -1	変数 **16**⇒変数 **15** の変化率　（変数 **15**－変数 **16**）÷変数 **16**
18	≥ 0	投稿pに周囲が付与した拡散数
19	≥ 0	投稿pの投稿者の過去の投稿に，周囲が付与してきた平均拡散数
20	≥ -1	変数 **19**⇒変数 **18** の変化率　（変数 **18**－変数 **19**）÷変数 **19**

表 3 各説明変数の重要度

	Average $(N=8)$	**Cluster 1** $(N=4)$	**Cluster 2** $(N=4)$
1	0.817	0.558	1.076
2	1.408	0.400	2.415
3	3.983	2.980	4.985
4	1.668	0.207	3.129
5	38.498	**54.227**	**22.768**
6	0.454	0.092	0.815
7	0.035	0.043	0.026
8	0.637	0.564	0.710
9	0.440	0.137	0.742
10	5.185	3.394	6.977
11	0.509	0.389	0.630
12	4.170	4.985	3.355
13	0.318	0.139	0.496
14	3.359	2.523	4.196
15	16.929	11.528	**22.331**
16	4.688	2.730	6.646
17	11.597	10.869	12.324
18	1.297	1.583	1.012
19	2.723	1.803	3.642
20	1.286	0.847	1.725

第 2 クラスタでは，変数 5 だけでなく「その投稿に周囲がどの程度お気に入り登録を行ったか（変数 15）」も同程度に重要視されている．ここに該当する 4 名は「投稿者が誰であるか」の他に「その投稿に周囲が肯定的に反応しているか」が主な要因で，その投稿にお気に入り登録を行う人物らであると解釈できる．

平均については，変数 5，変数 15，変数 17 など，お気に入り登録が関与する変数の重要度が高くなっており，直感と一致する結果となった．

4 おわりに

追加実験では，被験者数を増やす事で，ユーザ群を 2 クラスタ以上にタイプ分類できる事も分かっており，これは後発の論文[4]で議論する．

参考文献

[1] T. Chen, et.al., “XGBoost: A Scalable Tree Boosting System,” KDD’16, pp.785–794 (2016).

[2] G. Ke, et.al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” NIPS’17, pp.3149–3157 (2017).

[3] A. V. Dorogush, et.al., “CatBoost: Gradient Boosting with Categorical Features Support,”, arXiv:1810.11363 (2018).

[4] 荒澤, et.al., “投稿に付与されるお気に入り登録数を予測するための教師あり学習に基づく SNS ユーザのライフログデータ解析”, IEICE-LIOS (2023 年 3 月予定).