C-01

機械学習を用いた YouTube チャンネル登録者数の予測

Predicting YouTube Channel Subscribers Using Machine Learning

松清 綾大†　　伊藤 淳子†　　宗森 純†
Ryota Matsukiyo　　Junko Itou　　Jun Munemori

1. はじめに

近年，YouTube[1]上に自身の動画配信チャンネルを開設し，オリジナルの動画や音楽コンテンツを制作しアップロードする YouTuber と呼ばれる個人や団体が多く存在する．YouTube には，チャンネル登録という機能があり，視聴者は好きな YouTube チャンネルを登録することでそのチャンネルに関する通知などを受け取ることができる．1 つのチャンネルに対し何人がチャンネル登録をしているかを表すチャンネル登録者数は，YouTuber の人気や認知度を示す指標としてたびたび用いられる．また，人気があり，チャンネル登録者数が多い YouTuber は視聴者の生活習慣や消費行動に大きな影響力を持つため， YouTuber を起用したマーケティングを行う企業なども存在し，その数は増加している[2]．このことから，YouTube チャンネル登録者数の予測は様々な面で有用であると考えられる．

本研究ではオープンデータである YouTube Data API[3]から取得したデータでデータセットを作成し，特徴量の相関係数を求め，チャンネル登録者数の増減傾向を分析する．また，教師あり機械学習法の一つである線形回帰を使用することで，1 か月後のチャンネル登録者数と 1 か月間のチャンネル登録者増加数の予測モデルを作成し，その精度を検討する．

2. 関連研究

田中らは，YouTube にアップロードされた動画に対し，アップロード初期の視聴数の推移パターンと視聴数の絶対値を用い，教師あり機械学習法の一つである単純ベイズ分類器を適用することで，長期間にわたり高人気を維持する動画を予測し，その精度を評価した[4]．田中らはこれにより，初期の視聴数の絶対値のみで予測する場合より，高精度で将来にわたって高人気を維持する動画を予測できることを明らかにした．高人気を維持する動画の予測は，メディアの広告の配置など様々な面で有用ではあるが，企業が YouTuber を起用したマーケティングを行う際は，広告塔となる YouTube チャンネル自体の人気や認知度を示すチャンネル登録者数の予測がより有用だと考えられる．

3. データセットの作成と分析

YouTube Data API 利用し，機械学習に用いるデータセットを作成した．

3.1 取得データ

YouTube Data API から， 2020 年 4 月までに開設され，かつデータが取得可能であった海外を含む 81265 チャンネルに対し，2020 年 6 月 10 日，2020 年 6 月 17 日，2020 年 7 月 17 日の

・チャンネル登録者数
・投稿動画数
・動画再生回数

を取得し，データセットを作成した．作成したデータセットについて，特徴量ごとに総和をチャンネル数で割った平均値を表 1 に示す．

表 1:作成したデータセットのチャンネルごとの平均値

特徴量	平均値
チャンネル登録者数(2020 年 6 月 10 日)	354034.8
チャンネル登録者数(2020 年 6 月 17 日)	355989.1
チャンネル登録者数(2020 年 7 月 17 日)	363790.2
投稿動画数(2020 年 6 月 10 日)	988.5
投稿動画数(2020 年 6 月 17 日)	991.3
投稿動画数(2020 年 7 月 17 日)	1003.6
動画再生回数(2020 年 6 月 10 日)	120130000.0
動画再生回数(2020 年 6 月 17 日)	120836100.0
動画再生回数(2020 年 7 月 17 日)	123804700.0

3.2 YouTube チャンネル登録者数の詳細

YouTube Data API で取得可能なチャンネル登録者数の値はその数によって省略されている[5]．チャンネル登録者数が 4 桁のチャンネルはチャンネル登録者数が 1 人変化すると公開されるチャンネル登録者数が更新され，チャンネル登録者数が 5 桁のチャンネルはチャンネル登録者数が 10 人変化するとチャンネル登録者数が更新される．このように，登録者数に応じて公開されるチャンネル登録者数の省略方法が異なる．チャンネル登録者数が更新される単位については表 2 に示す．

表 2: チャンネル登録者数が更新される単位

チャンネル登録者数(人)	チャンネル登録者数が更新される単位(人)
1,000	1
1,000～9,999	10
10,000～99,999	100
100,000～999,999	1,000
1,000,000～9999,999	10,000
10,000,000～99,999,999	100,000
100,000,000～999,999,999	1,000,000

4. チャンネル登録者数の予測と考察

作成したデータセットの特徴量を使用し，教師あり機械学習の一つである線形回帰によりチャンネル登録者数の予測を行う．

4.1 説明変数と目的変数

説明変数として，2020 年 6 月 10 日と 2020 年 6 月 17 日のチャンネル登録者数，投稿動画数，動画再生回数と 2020 年 6 月 10 日～2020 年 6 月 17 日のチャンネル登録者増加数，投稿動画数の増加数，動画再生回数の増加数の計 9 つを使

† 和歌山大学，Wakayama University

用する．目的変数は 2020 年 7 月 17 日のチャンネル登録者とする．また，説明変数と目的変数の相関係数を求めた（表 3）．2020 年 6 月 10 日のチャンネル登録者数と 2020 年 7 月 17 日のチャンネル登録者数の相関係数が 0.9997 であり，2020 年 6 月 17 日のチャンネル登録者数と 2020 年 7 月 17 日のチャンネル登録者数の相関係数が 0.9998 となっている．

表 3: 2020 年 7 月 17 日のチャンネル登録者数との相関係数

説明変数	相関係数
チャンネル登録者数(2020 年 6 月 10 日)	0.9997
チャンネル登録者数(2020 年 6 月 17 日)	0.9998
投稿動画数(2020 年 6 月 10 日)	0.1195
投稿動画数(2020 年 6 月 17 日)	0.1194
動画再生回数(2020 年 6 月 10 日)	0.8085
動画再生回数 (2020 年 6 月 17 日)	0.8086
チャンネル登録者増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.6750
投稿動画数の増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.0103
動画再生回数の増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.3440

4.2 予測モデルの作成とその精度

線形回帰による予測モデルの作成には Python の機械学習ライブラリ scikit-learn[6]を使用する．81265 チャンネルのデータのうち，75%を訓練セット，25%をテストセットにランダムに割り当てた．訓練セットを利用し，4.1 節で示した 9 つの説明変数のうち，1 つのみ利用した予測モデルを 9 つ作成し，訓練セットとテストセットに対する決定係数を求め,その精度を示した（表 4）．

表 4: 1 つの説明変数を利用した 2020 年 7 月 17 日のチャンネル登録者数を予測する線形モデルの精度

説明変数	訓練セットスコア	テストセットスコア
チャンネル登録者数 (2020 年 6 月 10 日)	0.9994	0.9995
チャンネル登録者数 (2020 年 6 月 17 日)	0.9996	0.9996
投稿動画数 (2020 年 6 月 10 日)	0.0168	0.0057
投稿動画数 (2020 年 6 月 17 日)	0.0168	0.0056
動画再生回数 (2020 年 6 月 10 日)	0.7056	0.4902
動画再生回数 (2020 年 6 月 17 日)	0.7058	0.4903
チャンネル登録者増加数 (2020 年 6 月 10 日～ 2020 年 6 月 17 日)	0.4557	0.4513
投稿動画数の増加数 (2020 年 6 月 10 日～ 2020 年 6 月 17 日)	0.1065	0.1447
動画再生回数の増加数 (2020 年 6 月 10 日～ 2020 年 6 月 17 日)	0.0001	0.0001

4.3 考察

2020 年 6 月 10 日のチャンネル登録者数と 2020 年 7 月 17 日のチャンネル登録者数の相関係数と 2020 年 6 月 17 日のチャンネル登録者数と 2020 年 7 月 17 日のチャンネル登録者数の相関係数が非常に高いが，海外の YouTube チャンネルにはチャンネル登録者数が 1 億人を超えるような，チャンネル登録者数が非常に多いチャンネルがいくつか存在するため，その影響で相関係数が 1 に近い値になった可能性があると考えられ，説明変数の考察をするにあたり，別の手法を検討する必要があると考えられる．また，このことから，チャンネル登録者数を用いた予測モデルの精度がほぼ 1 ではあることが，チャンネル登録者数の予測モデルが良い予測モデルであることを示しているわけではないと考えられる．

5. チャンネル登録者増加数の予測と考察

次に，4 章と同じ手法でチャンネル登録者増加数の予測を行った．

5.1 説明変数と目的変数

説明変数として 4.1 節で示したものと同様の 9 つを使用する．目的変数は 2020 年 6 月 17 日から 2020 年 7 月 17 日までの 1 か月間のチャンネル登録者増加数とする．また，説明変数と目的変数の相関係数は表 5 に示す．

表 5: 2020 年 6 月 17 日～ 2020 年 7 月 17 日の チャンネル登録者増加数との相関係数

説明変数	相関係数
チャンネル登録者数(2020 年 6 月 10 日)	0.7025
チャンネル登録者数(2020 年 6 月 17 日)	0.7051
投稿動画数(2020 年 6 月 10 日)	0.1173
投稿動画数(2020 年 6 月 17 日)	0.1174
動画再生回数 (2020 年 6 月 10 日)	0.6299
動画再生回数(2020 年 6 月 17 日)	0.6318
チャンネル登録者増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.8019
投稿動画数の増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.3743
動画再生回数の増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.0220

5.2 予測モデルの作成とその精度

4.2 節と同様の手法で，9 つの説明変数のうち 1 つのみ利用した予測モデルを 9 つ作成し，訓練セットとモデルセットに対する決定係数を求め，その精度を示した（表 6）．2020 年 6 月 10 日～2020 年 6 月 17 日のチャンネル登録者増加数を説明変数とした予測モデルの精度が最も高い結果となった．また，比較的相関係数が高い，2020 年 6 月 10 日～2020 年 6 月 17 日のチャンネル登録者増加数，2020 年 6 月 10 日のチャンネル登録者数，2020 年 6 月 17 日のチャンネル登録者数の 3 つの説明変数を利用した予測モデルと 9 つ全ての説明変数を利用した予測モデルを作成し，その精度を求めた（表 7）．9 つ全ての説明変数を利用した予測モデルではテストセットスコアが 0.7229 とやや高い精度となった．

表 6: 1 つの説明変数を利用した 2020 年 6 月 17 日～2020 年 7 月 17 日のチャンネル登録者増加数を予測する線形モデルの精度

説明変数	訓練セットスコア	テストセットスコア
チャンネル登録者数 (2020 年 6 月 10 日)	0.4930	0.4949
チャンネル登録者数 (2020 年 6 月 17 日)	0.4968	0.4982
投稿動画数 (2020 年 6 月 10 日)	0.0151	0.0088
投稿動画数 (2020 年 6 月 17 日)	0.0151	0.0088
動画再生回数 (2020 年 6 月 10 日)	0.3848	0.4375
動画再生回数 (2020 年 6 月 17 日)	0.3874	0.4394
チャンネル登録者増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.6343	0.6694
投稿動画数の増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.1065	0.1447
動画再生回数の増加数 (2020 年 6 月 10 日～2020 年 6 月 17 日)	0.0005	0.0004

表 7: 複数の説明変数を利用した 2020 年 6 月 17 日～2020 年 7 月 17 日のチャンネル登録者増加数を予測する線形モデルの精度

説明変数	訓練セットスコア	テストセットスコア
比較的相関係数が高い 3 つの説明変数	0.6890	0.7169
9 つ全ての説明変数	0.6961	0.7229

5.3 考察

2020 年 6 月 10 日～2020 年 6 月 17 日のチャンネル登録者増加数と 2020 年 6 月 17 日～ 2020 年 7 月 17 日のチャンネル登録者増加数の相関係数が 0.8019 と高く，2020 年 6 月 10 日～2020 年 6 月 17 日のチャンネル登録者増加数を説明変数として利用した場合の予測モデルの精度が訓練セットスコアとテストセットスコアともにやや高かったことから，過去 1 週間チャンネル登録者数が増加したチャンネルはその後 1 か月間も同じようにチャンネル登録者数が増加する傾向にあると考えられる．

また，9 つ全ての説明変数を利用した予測モデルのテストセットスコアが 0.7229 となり，説明変数を 1 つ利用したどの予測モデルよりも良いテストセットスコアとなった．線形モデルは多くの説明変数を利用することで過剰適合となる可能性が高くなるが，今回は訓練セットスコアとテストセットスコアが近い値であるため，複数の説明変数を利用することで，より良い予測モデルを作成することが可能となった可能性がある．

6. まとめ

本研究では，説明変数として YouTube Data API から比較的取得が容易なオープンデータを使用し，教師あり機械学習法の一つである線形回帰を用いて，1 か月後のチャンネル登録者数と 1 か月間のチャンネル登録者増加数の予測を行った．しかし，突出してチャンネル登録者数が多いチャンネルがいくつか存在するため，相関係数による説明変数の考察や決定係数による予測モデルの評価では不十分であり，他の手法を検討する必要があると考えられる．

今後の展望として，YouTube Data API から取得可能なデータだけでなく，YouTube と関連がある SNS などから，より多くの特徴量を使用し，クラス分類による機械学習を行うことで，突出してチャンネル登録者数が増加している YouTube チャンネルがどのような特徴があるかの分析を検討する．

参考文献

[1]YouTube “https://www.youtube.com/” (2020.7.21 確認).
[2]鎌田 和樹: 講座開講記念講演「ネット動画マーケティングから見るクリエイター育成とビジネス開発の可能性」,DHU JOURNAL Vol.04 2017 - Daily Life with Super Technologies- (2017).
[3] YouTube Data API “https://developers.google.com/youtube/v3” (2020.7.21 確認).
[4]田中 達也, 村田 正幸: ユーザー生成コンテンツの視聴数推移パターン分析と人気推移予測, 信学技報, vol.116, no.137, IN2016-31, pp.49-54 (2016).
[5]チャンネル登録者数 “https://support.google.com/youtube/answer/6051134” (2020).
[6] scikit-learn “https://scikit-learn.org/stable/“ (2020.7.21 確認).