「マルチメディア，分散，協調とモバイル
(DICOMO2019) シンポジウム」 令和元年7月

# 動画視聴ログを用いた深層学習による視聴者の年代推定

永田 尚志[1]

**概要**：動画配信サービスのパーソナライズ化を指向し、大量の動画視聴ログを元に深層学習を用いて動画視聴者の年代推定を行ったので報告する。投稿動画の配信サービスを対象とし、動画視聴ログに含まれる視聴者の年代ごとの視聴時間、視聴曜日、人気の動画などの傾向に着目し、深層学習への入力データの表現方法、教師データの効果的な学習手法を提案する。投稿動画の人気度や視聴時間の分割数などを考慮した複数の入力データ表現を定義し、学習と推定を行った結果、最大で約 15%の精度向上を実現する入力データ表現を明らかにした。

# Age estimation for viewers by using Deep Learning based on video viewing logs

HISASHI NAGATA[1]

## 1. はじめに

近年、動画共有サービスが普及し、年齢層を問わず多数の動画が視聴されている。動画視聴傾向には視聴者ごとに趣味嗜好が現れるため、視聴動画のジャンルによる動画のレコメンドは容易だが、年代に応じたマーケティングには結びつけることは難しい。また、技術面においては分散処理技術と深層学習の発展は目覚ましく、蓄積されたビッグデータから高度な解析が可能となり新たな付加価値が創出されている。

本検討では、一般の動画共有サービスを対象とし、「視聴者がどの投稿者の動画をいつ視聴したか」という動画視聴ログを元に視聴者の年齢を推定し、年代に基づくパーソナルサービスに結びつけることを指向する。動画視聴ログには「年代ごとの教師データ量の不均衡」と「年代によらず類似性のある動画視聴傾向」と「特定の投稿者への視聴集中」と「同一動画投稿者の連続視聴」の特徴がある。これらの動画視聴ログの独特の特徴に着目した深層学習における適切な入力データ表現方法と教師データの効果的な学習手法の提案をし、視聴者の年代推定に対する精度を考察する。

[1] 西日本電信電話株式会社
NIPPON TELEGRAPH AND TELEPHONE WEST CORPORATION

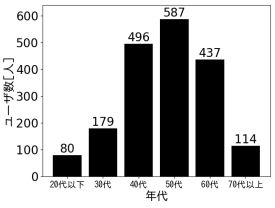


**図 1** 年代ごとのユーザ数

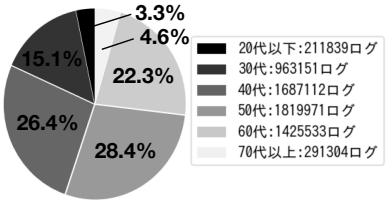


**図 2** ラベルごとのログ数の割合

| 動画投稿者 | 視聴日時 | 年代 | 視聴者名 |
|---|---|---|---|
| 動画投稿者 A | 2019 年 03 月 28 日 19 時 | 20 代以下 | 視聴者 1 |
| 動画投稿者 B | 2019 時 04 月 25 日 21 時 | なし | 視聴者 2 |
| 動画投稿者 A | 2019 年 04 月 28 日 21 時 | 20 代以下 | 視聴者 1 |
| 動画投稿者 C | 2019 年 05 月 01 日 09 時 | 40 代 | 視聴者 3 |
| 動画投稿者 C | 2019 年 05 月 01 日 10 時 | 70 代以上 | 視聴者 4 |

**表 1** 蓄積された動画視聴ログのデータ形式

## 2. 動画視聴ログの特徴

### 2.1 動画視聴ログのフォーマット

動画視聴ログのデータ形式は表 1 の通りである。動画視聴ログの項目には「動画投稿者」「視聴日時」「年代」「視聴者名」が記録される。動画投稿者は視聴された動画の投稿者名、視聴日時は動画が視聴された日時 (YYYYMMDDHH 形式)、視聴者は視聴した人を識別する固有名が記載される。年代については、1893 人の視聴者にアンケートを実

施し、回答内容をもとにラベルを付与して教師データとする。教師データから作成した学習モデルにより、日々蓄積しているラベルなしの動画視聴ログから年代推定を可能にする。今回は「20 代以下」「30 代」「40 代」「50 代」「60 代」「70 代以上」の 6 つの中からラベルを付与し、アンケートをしていない視聴者はラベルなしとする。表 1 の 1 行を 1 つの動画視聴ログとして定義し、ある一定期間内にサーバに蓄積されている約 650 万件の動画視聴ログを教師データとして用いる。また、半教師あり学習で用いるラベルなしの動画視聴ログを約 1000 万件利用した。

### 2.2 年代ごとの教師データ量の不均衡

各年代の教師データ量の差が問題となる。アンケートを実施した視聴者の年代ごとの内訳は図 1 の通りである。もっとも少ない「20 代以下」が 80 人、もっとも多い「50 代」が 587 人であり、年代ごとの視聴者数に最大 7.3 倍の開きがある。また、各年代の動画視聴回数は図 2 の通りである。もっとも少ない「20 代以下」が 211839 回、もっとも多い「50 代」で 1819971 回の動画が視聴されており、最大で動画視聴回数に関して 8.6 倍の開きがある。図 1 と図 2 を比較すると、視聴者数と視聴された動画視聴ログ数はおおよそ比例関係にあり、双方とも少ない順に「20 代以下」「70 代以上」「30 代」「60 代」「40 代」「50 代」となっている。

以上により、各年代の教師データ量の調整や学習モデルに対して教師データ量の少ない年代を学習させるために損失関数の調整を行う必要がある。

### 2.3 年代によらず類似性のある動画視聴傾向

教師データを集計すると、視聴される動画投稿者や視聴曜日や視聴時間帯に年代ごとに明らかな相違はなく、似たような視聴傾向が問題となる。各年代の人気動画投稿者上位 10 位以内を挙げると表 2 の通りである。人気動画投稿者は年代によらず、どの年代でも視聴される傾向がある。特に「40 代」の人気動画投稿者の大半は他年代でも人気となっている。さらに動画投稿者に着目すると、動画投稿者 E は全ての年代で人気となっている。その一方で、「20 代以下」と「70 代以上」では人気動画投稿者が他年代で人気となることが少なく、異なる視聴傾向が見られる。よって、一定数 ($m \in \mathbb{N}$) の動画視聴ログを考慮することで、視聴した動画投稿者の組合せや動画投稿者に対する視聴回数を各年代の特徴として学習させる。

各年代の月火水木金土日の曜日ごとの視聴割合は表 3 の通りである。「40 代」を除き、年代によらず土日に多くの動画視聴がされる傾向がある。特に「30 代」では平日の視聴が少なく土日の視聴が極端に多いことが分かる。その一方で、「20 代以下」と「40 代」では平日に視聴が多いという特徴が見られる。これらの特徴は学校や仕事などの生活リズムが影響していると考えられる。

| 投稿者＼年代 | 20 代以下 | 30 代 | 40 代 | 50 代 | 60 代 | 70 代以下 |
|---|---|---|---|---|---|---|
| 1 位 | A | A | O | E | E | Q |
| 2 位 | B | K | E | G | A | E |
| 3 位 | C | F | A | P | G | Z |
| 4 位 | D | L | G | S | P | G |
| 5 位 | E | E | P | T | Q | P |
| 6 位 | F | M | F | K | W | a |
| 7 位 | G | N | L | U | X | b |
| 8 位 | H | B | B | Q | V | Y |
| 9 位 | I | G | Q | V | K | V |
| 10 位 | J | O | R | F | Y | c |

**表 2** 年代ごとの人気動画投稿者ランキング (大小のアルファベットは動画投稿者の固有名を表し、複数の年代で上位 10 以内となる動画投稿者をグレーで色分けする。)

| 曜日＼年代 | 20 代以下 | 30 代 | 40 代 | 50 代 | 60 代 | 70 代以上 |
|---|---|---|---|---|---|---|
| 月曜日 | 13.3 | 10.5 | 15.6 | 14.7 | 14.0 | 14.4 |
| 火曜日 | 14.1 | 13.1 | 14.5 | 13.0 | 13.5 | 13.7 |
| 水曜日 | 13.0 | 11.5 | 14.6 | 13.6 | 13.7 | 13.5 |
| 木曜日 | 15.0 | 11.0 | 14.4 | 13.7 | 13.7 | 13.3 |
| 金曜日 | 14.3 | 12.9 | 13.2 | 13.9 | 13.8 | 13.5 |
| 土曜日 | 18.1 | 17.6 | 13.7 | 15.7 | 15.8 | 15.7 |
| 日曜日 | 12.2 | 23.4 | 14.0 | 15.4 | 15.4 | 15.8 |

**表 3** 年代ごとの視聴曜日の割合 [%]（各年代で頻繁に視聴されている上位 2 つの曜日をグレーで色分けする。)

| 時間帯＼年代 | 20 代以下 | 30 代 | 40 代 | 50 代 | 60 代 | 70 代以上 |
|---|---|---|---|---|---|---|
| 0〜1 時 | 2.2 | **1.4** | 5.4 | 4.1 | 4.2 | 4.7 |
| 1〜2 時 | 2.1 | **1.0** | 3.7 | 3.8 | 3.4 | 3.8 |
| 2〜3 時 | **1.9** | **0.8** | 3.2 | 3.5 | 2.9 | 3.2 |
| 3〜4 時 | **1.9** | **0.7** | 2.8 | 3.0 | 2.7 | 3.0 |
| 4〜5 時 | **2.0** | **0.7** | 2.7 | 2.7 | 2.5 | 2.8 |
| 5〜6 時 | 2.5 | **0.9** | 2.9 | 2.5 | 2.5 | 2.6 |
| 6〜7 時 | 3.2 | **0.8** | 2.7 | 2.6 | 2.9 | 2.7 |
| 7〜8 時 | 5.0 | **1.0** | 2.7 | 2.9 | 3.2 | 3.2 |
| 8〜9 時 | 5.4 | 2.1 | 3.1 | 3.5 | 3.4 | 3.5 |
| 9〜10 時 | 4.2 | 2.9 | 3.4 | 3.9 | 3.7 | 3.6 |
| 10〜11 時 | 4.4 | 3.8 | 3.5 | 4.0 | 3.9 | 3.9 |
| 11〜12 時 | 4.9 | 3.3 | 3.7 | 4.0 | 4.2 | 4.0 |
| 12〜13 時 | 4.9 | 3.6 | 3.9 | 4.0 | 3.9 | 3.8 |
| 13〜14 時 | 4.9 | 3.9 | 4.4 | 4.5 | 4.1 | 4.1 |
| 14〜15 時 | 5.1 | 5.0 | 4.1 | 4.7 | 4.2 | 4.3 |
| 15〜16 時 | 5.3 | 5.1 | 4.2 | 4.7 | 4.6 | 4.5 |
| 16〜17 時 | 5.5 | 5.0 | 5.7 | 4.9 | 5.1 | 4.8 |
| 17〜18 時 | 6.2 | 8.3 | 5.6 | 5.3 | 5.9 | 4.9 |
| 18〜19 時 | 5.9 | 12.5 | 3.9 | 6.1 | 5.9 | 4.8 |
| 19〜20 時 | 6.4 | 11.9 | 4.8 | 5.6 | 5.5 | 4.8 |
| 20〜21 時 | 6.7 | 12.2 | 5.3 | 5.1 | 5.8 | 5.7 |
| 21〜22 時 | 4.2 | 6.0 | 6.4 | 5.0 | 5.7 | 6.4 |
| 22〜23 時 | 2.8 | 3.9 | 6.1 | 4.8 | 5.1 | 5.6 |
| 23〜0 時 | 2.4 | 3.2 | 5.9 | 4.5 | 4.9 | 5.2 |

**表 4** 年代ごとの視聴時間帯の割合 [%]（頻繁に視聴されている時間帯の上位 3 つをグレーで色分けし、視聴が少ない 2.0%以下の時間帯を太文字とする。）

各年代の 1 時間刻みの時間帯ごとの視聴割合は表 4 の通りである。テレビで高視聴率となるゴールデンタイムの時間帯付近で動画視聴も同様の傾向があり、年代によらず 17

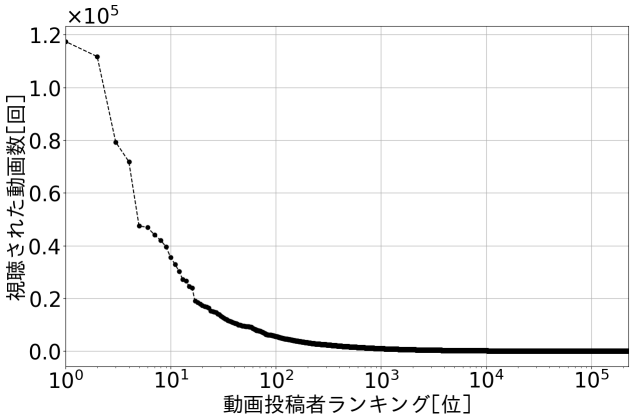


**図 3** ユーザの人気動画投稿者への視聴偏り

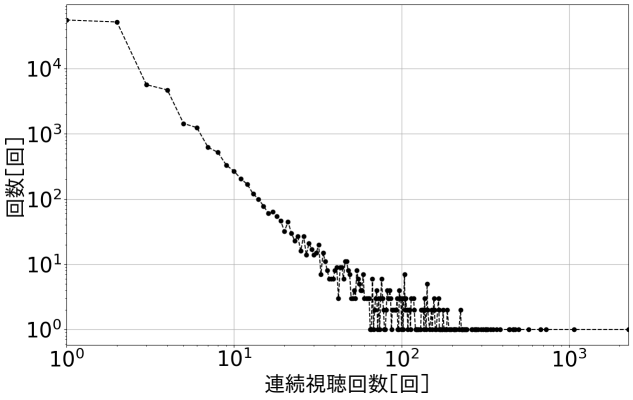


**図 4** 同一動画投稿者の連続視聴回数

時〜22 時の時間帯で動画視聴が頻繁にされている。特に「30 代」ではこの 5 時間で全体の 50.9%の視聴がされている。その一方で、「30 代」は深夜から早朝にかけて視聴が極端に少なく、「40 代」と「70 代以上」は 21〜1 時の時間帯の視聴が多く、「20 代以下」は昼に安定して視聴されるという各年代の特徴が見られる。視聴曜日と同様に、視聴時間帯は各年代の生活リズムが影響していると考えられる。

以上により、人気動画投稿者と視聴曜日と視聴時間帯の各々で、年代を問わず類似性が高い特徴が存在するが、ある程度の年代ごとの特徴も見られる。よって、複数の動画視聴ログからこれら 3 つの要素を適切に組合せることで年代ごとの特徴を効果的に学習させ、高精度に年代推定を実現することを目指す。

### 2.4 特定の投稿者への視聴集中

一部の動画投稿者に視聴が集中するという視聴集中が問題となる。動画投稿者ごとに視聴数に基づきランキング順でまとめたものが図 3 である。視聴された動画投稿者数は約 20 万人いるが、1 番視聴された動画投稿者は約 12 万回視聴されている。その一方で、1000 番目以降になると視聴回数は急激に減少し、約 1000 回しか視聴されていない。特定の動画投稿者に視聴が圧倒的に集中しているため、上位 1000 位以降は人気動画投稿者と比較してほとんど視聴されていないことが分かる。よって、不人気な動画投稿者は多数の視聴者に視聴されなく、個人特定には必要であるが年代推定には有効でないと考えられる。

以上により、動画投稿者は無数存在するが人気投稿者のみを学習対象とすることで効果的に学習を行う。そのため、視聴数による人気動画投稿者ランキングを作成し、上位 $n \in \mathbb{N}$ 人を学習対象とする。

### 2.5 同一の動画投稿者の連続視聴

同じ動画投稿者を連続視聴することが問題となる。同一の動画投稿者に対する連続視聴回数は図 4 の通りである。連続視聴回数の平均は 2.73 回である。また、連続視聴回数が 10 回以上となることは多々あり、最大で 2260 回連続視聴されるケースも存在する。これは動画共有サービスの動画は 1 本当たりの動画再生時間が短いという性質から、視聴者は気軽に自身の趣味嗜好に合った動画投稿者を連続視聴することに起因すると考えられる。そのため、視聴者ごとに時系列順に $m$ ログ抽出する方法があるが、同一投稿者の連続視聴による偏り解消のため、ランダムに $m$ ログ抽出することが有効であると考えられる。

## 3. 提案手法

### 3.1 入力データ

動画視聴ログの特徴に着目し、人気投稿者の視聴回数をベクトルで表現した入力データ形式とする。入力データは $\boldsymbol{x} = (x_1, x_2, \cdots, x_{n+1})$ の $n+1$ 次元ベクトル空間 $(\sum_{i=1}^{n+1} x_i = m)$ で表現する。$x_i \geq 0$ は $i \leq n$ 位の人気投稿者、$x_{n+1}$ はそれ以外の視聴回数とする。

動画投稿者に加えて視聴曜日を考慮するとき、視聴曜日の分割数を $l \in \{1, 2, 7\}$ で表すと、$\boldsymbol{x} = (x_1, x_2, \cdots, x_n, x_{n+1}, \cdots, x_{l(n+1)-1}, x_{l(n+1)})$ の $l(n+1)$ 次元ベクトル空間で表現する。$l = 2$ のとき、視聴曜日を「平日」「休日」で 2 分割にラベル付けし、$l = 7$ のとき、視聴曜日を「月曜」「火曜」「水曜」「木曜」「金曜」「土曜」「日曜」の 7 分割にラベル付けする。$x_i$ は $\lceil i/(n+1) \rceil$ の曜日に視聴した $i \bmod (n+1)$ 位の人気投稿者の視聴回数とする。$l = 1$ は視聴曜日の分割数が 1 であり、視聴曜日を入力データに含まないとする。

同様に動画投稿者に加えて視聴時間帯を考慮するとき、視聴時間帯の分割数を $k \in \{1, 2, 4, 6, 8\}$ と表すと、$\boldsymbol{x} = (x_1, x_2, \cdots, x_n, x_{n+1}, \cdots, x_{k(n+1)-1}, x_{k(n+1)})$ の $k(n+1)$ 次元ベクトル空間で表現する。例えば $k = 2$ のとき、視聴時間帯を「0〜11 時」「12〜23 時」で 2 分割でラベル付けする。$x_i$ は $\lceil i/(n+1) \rceil$ の時間帯に視聴した $i \bmod (n+1)$ 位の人気投稿者の視聴回数とする。$k = 1$ は視聴時間帯の分割数が 1 であり、視聴時間帯を入力データ形式に含まないとする。

さらに動画投稿者に加えて視聴曜日と視聴時間帯を考慮するとき、$k$ と $l$ を用いて $\boldsymbol{x} = (x_1, x_2, \cdots, x_n, x_{n+1}, \cdots,$

$x_{lk(n+1)-1}, x_{lk(n+1)})$ の $lk(n+1)$ 次元ベクトル空間で表現する。

1 つの入力データの中に異なる年代、異なる視聴者の動画視聴ログは含まない。また、1 つの入力データとして集計する動画視聴ログが $m$ ログ以上蓄積されていない視聴者に対しては、入力データを作成しない。視聴者ごとにランダムに $m$ ログずつ抽出して入力データを作成し、$m$ ログ抽出できなくなるまでこれを繰り返す。視聴者単位で入力データを教師データとテストデータに分別するため、同じ視聴者の入力データが教師データとテストデータの両方に含まれることはない。また、今回の評価では教師データとテストデータに分ける視聴者数の比率を 7:3 とする。

### 3.2 学習モデル

学習モデルは多層パーセプトロン、最適化は Adam、活性化関数は ReLU、ドロップアウト率 [1] は入力層 20%と中間層 50%、バッチサイズ [2] は 64 とする。

年代ごとの教師データ量不均衡性を解消させるため、教師データ量と損失関数による調整を行う。教師データ量については、オーバーサンプリング手法である SMOTE[3] と、アンダーサンプリング手法である NearMiss[4] を適用する。損失関数については交差エントロピーを用いてデータ量に応じた重み付けを行う。$N$ を教師データ数、$d_{\alpha\beta}$ を $\alpha$ 番目の教師データの正解ラベル $\beta$ のみ 1 でそれ以外 0、$p_{\alpha\beta}$ を $\alpha$ 番目の教師データに対する正解ラベル $\beta$ の推定確率、$a_\beta \in \mathbb{N}$ は各年代のデータ量とすると、重み付けされた損失関数 $E_\gamma$ は式 1 で表される。教師データ量に反比例、二乗に反比例させるときは各々、$\gamma = 1$、$\gamma = 2$ となる。

$$E_\gamma = -\frac{1}{N}\sum_{\alpha=1}^{N}\sum_{\beta=1}^{6}\frac{1}{a_\beta^\gamma} d_{\alpha\beta}\log p_{\alpha\beta} \qquad (1)$$

また、推定精度向上のため、半教師あり学習の手法である self-training を用いて、ラベルなしデータから教師データを増やす。ラベルなしデータを学習モデルに入力し、推定確率が 99.9% 以上の高確率で推定されたデータに対して、推定された結果をラベルに付与して教師データに追加する。さらに過学習を防ぐために、L1 正則化と L2 正則化を適用する。$\mu \in \mathbb{N}$ 層目の $\nu \in \mathbb{N}$ ノードから $\omega+1$ 層目の $\nu \in \mathbb{N}$ ノードへの重みを $w_{\mu\nu}^{(\omega)}$、定数を $\lambda_1 \in \mathbb{R}$ とすると、L1 正則化項を含んだ損失関数 $L1$ は式 2 で表される。

$$L1 = -\frac{1}{N}\sum_{\alpha=1}^{N}\sum_{\beta=1}^{6} d_{\alpha\beta}\log p_{\alpha\beta} + \lambda_1 \sum_{\mu,\nu,\omega} |w_{\mu\nu}^{(\omega)}| \qquad (2)$$

同様に定数を $\lambda_2 \in \mathbb{R}$ とすると、L2 正則化項を含んだ損失関数 $L2$ は式 3 で表される。

$$L2 = -\frac{1}{N}\sum_{\alpha=1}^{N}\sum_{\beta=1}^{6} d_{\alpha\beta}\log p_{\alpha\beta} + \frac{\lambda_2}{2}\sum_{\mu,\nu,\omega} (w_{\mu\nu}^{(\omega)})^2 \qquad (3)$$

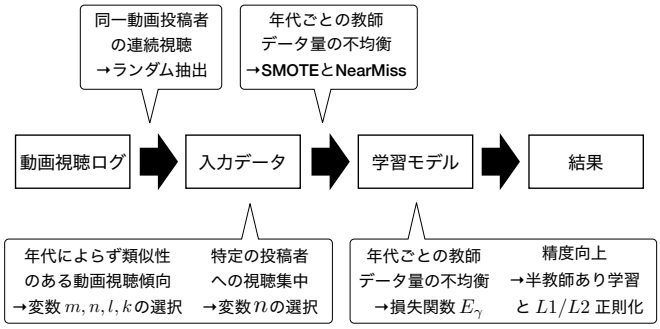


**図 5** 動画視聴ログの特徴に対する提案手法の全体像

### 3.3 提案手法の全体像

年代推定に関する提案手法の全体像は図 5 の通りである。同一動画投稿者の連続視聴の問題に対しては、動画視聴ログをランダムに抽出することで複数の動画投稿者の組合せを学習させる。特定の投稿者への視聴集中の問題に対しては、変数 $n$ の適切な値の選択することで効果的に年代ごとの特徴を学習させる。年代によらず類似性のある動画視聴傾向の問題に対しては、変数 $m,n,l,k$ の適切な値を選択することで年代ごとの特徴を明確にする。年代ごとの教師データ量の不均衡の問題に対しては、教師データ量と損失関数による調整を行い、少ない教師データを学習させる。また、半教師あり学習や L1/L2 正則化を用いることで更なる精度向上を図る。

## 4. 結果

### 4.1 評価方法

精度による評価はテストデータごとに算出する正解率 (正解数/テストデータ数) とマクロ適合率 (各年代の正解率の平均値) により行う。以降、マクロ適合率を適合率と呼び、正解率と適合率は 5 回計測した平均値で評価する。

### 4.2 $m$ の選択とランダム抽出

まず、1 つの入力データとして集計する動画視聴ログ数である $m$ について検討する。また、動画視聴ログを時系列順に入力データを作成する方法とランダムに選出して入力データを作成する方法との精度比較をする。学習対象とする人気投稿者数は上位 $n = 3000$ 人に固定し、$m$ の値を $100, 200, 300, 400, 500, 600, 700, 800, 900, 1000$ と変化させたときの結果は表 5 となる。$m$ を大きくすると時系列順とランダムの双方で正解率と適合率について上昇傾向にある。

考慮する動画視聴ログを増やすことで複数の動画投稿者が入力データに反映され、人気動画投稿者は各年代で重複はあるが多数の動画投稿者の組合せにより各年代の特徴を学習していると考えられる。また、時系列順と比較し、ランダムでは全ての $m$ に対して正解率と適合率は高くなっている。動画視聴ログをランダム抽出することにより、図

| m の値 (n=3000) | | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 時系列順 | 正解率 | 33.29 | 34.24 | 34.56 | 35.03 | 34.85 | 36.38 | 35.77 | 36.71 | 37.73 | 35.65 |
| | 適合率 | 24.83 | 26.44 | 26.47 | 25.70 | 25.89 | 27.18 | 27.79 | 30.18 | 27.63 | 28.20 |
| ランダム | 正解率 | 37.49 | 38.27 | 39.52 | 37.00 | 39.71 | 39.57 | 39.67 | 42.59 | 40.21 | 38.46 |
| | 適合率 | 30.29 | 30.09 | 33.35 | 30.54 | 33.98 | 34.55 | 34.30 | 35.85 | 33.85 | 36.22 |

**表 5** 入力する動画視聴ログ数 $m$ の変化による精度 [%]

| n の値 | 100 | 300 | 500 | 1000 | 3000 | 5000 | 10000 | 30000 | 50000 | 100000 |
|---|---|---|---|---|---|---|---|---|---|---|
| 正解率 | 35.16 | 34.88 | 37.24 | 36.63 | 39.71 | 38.23 | 36.24 | 37.09 | 36.18 | 37.27 |
| 適合率 | 29.72 | 28.64 | 29.43 | 31.06 | 33.98 | 30.69 | 30.35 | 31.13 | 31.43 | 33.79 |

**表 6** 投稿者数 n の変化による精度 [%]

| n の値 | | 1000 | 3000 | 5000 | 10000 |
|---|---|---|---|---|---|
| l=2, k=1 | 正解率 | 36.72 | 40.70 | 41.54 | 42.65 |
| | 適合率 | 28.18 | 32.86 | 33.55 | 33.54 |
| l=7, k=1 | 正解率 | 38.51 | 43.59 | 39.24 | 42.95 |
| | 適合率 | 32.47 | 34.56 | 34.87 | 32.53 |

**表 7** 視聴曜日 $l$ と人気動画投稿者数 $n$ の変化による精度 [%]

| n の値 | | 1000 | 3000 | 5000 | 10000 |
|---|---|---|---|---|---|
| l=1, k=2 | 正解率 | 37.56 | 41.16 | 39.92 | 42.12 |
| | 適合率 | 34.10 | 30.62 | 32.90 | 32.17 |
| l=1, k=4 | 正解率 | 38.93 | 42.52 | 43.54 | 47.90 |
| | 適合率 | 33.21 | 34.73 | 31.24 | 40.44 |
| l=1, k=6 | 正解率 | 38.52 | 40.41 | 41.89 | 46.62 |
| | 適合率 | 33.63 | 32.65 | 33.27 | 35.64 |
| l=1, k=8 | 正解率 | 35.61 | 39.74 | 43.27 | 45.33 |
| | 適合率 | 30.49 | 32.07 | 37.24 | 37.21 |

**表 8** 視聴時間帯 $k$ と人気動画投稿者数 $n$ の変化による精度 [%]

4 に挙げた連続視聴による偏りを緩和していると考えられる。$m$ の最適値については、$m$ を大きくすると精度向上する一方で、学習対象とする視聴者数と教師データ量が減少する。以上により、入力データは精度が高いランダム手法を用いて作成し、ある程度の精度と視聴者数と教師データ量を保つことが可能である $m = 500$ に固定し、以降の評価を行う。

### 4.3 $n$ の選択

続いて学習対象とする人気動画投稿者数である $n$ について検討する。$n$ の値を $100, 300, 500, 1000, 3000, 5000, 10000, 30000, 50000, 100000$ と変化させたときの結果は表 6 となる。$n$ を増加させると正解率と適合率は上昇傾向にあり $n = 3000$ で最大値となるが、それ以上増加させても低下する。図 3 で挙げた人気動画投稿者への視聴集中により、上位数千位までが多数の視聴者に視聴されるため分類に有効であるが、それ以降は少数の視聴者にしか視聴されていなく分類に不向きであると考えられる。以上により、$n = 1000, 3000, 5000, 10000$ に固定し、以降の評価を行う。

### 4.4 $l, k$ の選択

次に視聴曜日の分割数である $l$ について検討する。$l$ の値を $2, 7$、人気動画投稿者数 $n$ を $1000, 3000, 5000, 10000$ と変化させたときの結果は表 7 となる。視聴曜日を考慮することで正解率と適合率は上昇する。また、視聴曜日の分割数に着目すると $l$ を大きくすると精度向上し、視聴曜日を考慮しないとき ($l = 1$) は $n = 3000$ で精度が最大となり、それ以降 $n$ を増加させても精度向上しないが、視聴曜日を考慮することで $n = 10000$ まで精度向上する。

同様に視聴時間帯の分割数である $k$ について検討する。$k$ の値を $2, 4, 6, 8$、人気動画投稿者数 $n$ を $1000, 3000, 5000, 10000$ と変化させたときの結果は表 7 となる。視聴時間帯を考慮することで正解率と適合率は上昇する。また、視聴曜日の分割数に着目すると $k = 4$ のとき精度は最大となり、それ以上増加させても精度は減少する。視聴曜日と同じように視聴時間帯を考慮することで $n = 10000$ まで増加させても精度向上する。

さらに視聴曜日と視聴時間帯の組合せについて検討する。視聴曜日の分割数 $l$ については大きくすると精度向上したため $l = 2, 7$ と変化させ、視聴時間帯については $k = 4$ 以上増加させても精度向上しないので $k = 2, 4$ と変化させ、$n = 5000$ と固定したときの結果は表 9 となる。視聴曜日と視聴時間帯を考慮しても、視聴曜日のみ、視聴時間帯のみと比較して精度は低くなる。入力データに視聴曜日と視聴時間帯の 2 要素を追加することで、入力層の次元数が大幅に増加し過学習すると考えられる。

正解率と適合率が最大となるパラメータ $m, n, l, k$ の組合せは $m = 500, n = 10000, l = 1, k = 4$ である。このときの年代別の適合率の 5 回平均は表 10 の通りである。exact は推定年代とラベル年代が完全一致、1-off は推定年代がラベル年代の $\pm 1$ 年代以内となる正解率を表す。20 代以下と 70 代以下の正解率が極端に低い。これは図 2 の教師データ量の差によると考えられる。1-off を見ると 70 代以上の正解率が大幅に上昇している。60 代と 70 代以上で、表 2 より人気動画投稿者上位 10 人の 6 人が重複しており、さらに表 4 により 17〜1 時の夕方から深夜の時

| n の値 | | 5000 |
|---|---|---|
| l=2, k=2 | 正解率 | 40.31 |
| | 適合率 | 33.24 |
| l=2, k=4 | 正解率 | 40.62 |
| | 適合率 | 31.36 |
| l=7, k=2 | 正解率 | 39.54 |
| | 適合率 | 32.29 |
| l=7, k=4 | 正解率 | 39.65 |
| | 適合率 | 29.68 |

**表 9** 視聴曜日 $l$ と視聴時間帯 $k$ の変化による精度 [%]

| | exact | 1-off |
|---|---|---|
| 20 代以下 | 11.88 | 48.18 |
| 30 代 | 45.72 | 73.81 |
| 40 代 | 36.94 | 86.39 |
| 50 代 | 44.09 | 84.81 |
| 60 代 | 55.95 | 82.82 |
| 70 代以上 | 6.00 | 52.92 |
| 全年代 | 47.90 | 81.63 |

**表 10** 年代ごとの精度 [%]

| SMOTE | 正解率 | 45.23 |
|---|---|---|
| | 適合率 | 35.76 |
| NearMiss | 正解率 | 20.36 |
| | 適合率 | 18.82 |
| 損失関数 $E_1$ | 正解率 | 44.18 |
| | 適合率 | 34.55 |
| 損失関数 $E_2$ | 正解率 | 47.28 |
| | 適合率 | 36.98 |

**表 11** 教師データ量不均衡に有効な手法の精度 [%]

| | exact | 1-off |
|---|---|---|
| 20 代以下 | 10.99 | 33.78 |
| 30 代 | 40.62 | 80.19 |
| 40 代 | 46.17 | 89.57 |
| 50 代 | 53.59 | 86.46 |
| 60 代 | 53.96 | 79.76 |
| 70 代以上 | 10.16 | 46.56 |
| 全年代 | 47.28 | 81.76 |

**表 12** 年代ごとの精度 [%]

| self-training | 正解率 | 46.32 |
|---|---|---|
| | 適合率 | 36.96 |
| L1 正則化 $\lambda_1 = 0.0001$ | 正解率 | 45.95 |
| | 適合率 | 40.45 |
| L2 正則化 $\lambda_2 = 0.01$ | 正解率 | 48.17 |
| | 適合率 | 40.87 |

**表 13** 精度向上のための手法ごとの精度 [%]

| | exact | 1-off |
|---|---|---|
| 20 代以下 | 11.22 | 38.49 |
| 30 代 | 41.50 | 72.30 |
| 40 代 | 35.39 | 89.66 |
| 50 代 | 60.46 | 87.62 |
| 60 代 | 53.87 | 78.44 |
| 70 代以上 | 25.19 | 71.61 |
| 全年代 | 48.17 | 82.70 |

**表 14** 年代ごとの精度 [%]

間帯での視聴が多いという類似性の高い特徴によるためと考えられる。以上により、高精度に推定が可能である $m = 500, n = 10000, l = 1, k = 4$ に固定し、以降の評価を行う。

### 4.5 教師データ量と損失関数の調整

教師データ量と損失関数の調整を適用すると表 11 の結果が得られる。全ての手法に関して、正解率と適合率は減少している。SMOTE は似た教師データを増加することによる過学習、NearMiss は教師データ量の減少による学習不足、$E_1$ と $E_2$ は教師データ量の多い年代の正解率の減少によると考えられる。$E_2$ の exact と 1-off は表 14 の通りである。教師データ量の少ない年代が精度向上している。

### 4.6 精度向上手法の適用

最後に教師データ量を増加させる self-training と過学習に有効な L1 正則化と L2 正則化を適用すると表 13 の結果が得られる。self-training は正解率と適合率の差が大幅に広がっている。教師データ量の多い年代の教師データが更に増加していると考えられる。L1 正則化は正解率と適合率は減少するが、その差は減少している。L1 正則化の効果により、本来は分類に有効な入力層における $x_i$ の寄与を小さくしていると考えられる。L2 正則化は正解率と適合率が増加している。入力層における $x_i$ の極端に大きな寄与を無くし、過学習を軽減していると考えられる。L2 正則化を適用したときの exact と 1-off は表 14 の通りである。L2 正則化により、1-off についても精度向上し、特に「70 代以上」が学習されている。

## 5. まとめ

動画視聴ログを用いて深層学習による視聴者の年代推定を試みた。入力データ形式には人気動画投稿者と視聴曜日と視聴時間帯を用いた。人気動画投稿者数については、上位 $n$ 人を選出することで効果的に学習が行え、$n$ を増やすと精度が向上する傾向があるが、数千程度を越えるとそれ以上は精度が低下すると分かった。人気動画投稿者以外は複数の視聴者に視聴されないためと考えられる。また、動画視聴ログには各年代で人気動画投稿者、視聴曜日、視聴時間帯に関して、明確な特徴差は見られなかったが、それらを適切に組合せて入力データを構成することで各年代の視聴傾向が掴むことができた。組合せ方は情報量を単純に多くすれば精度向上するわけではなく、動画視聴ログによる年代推定においては $n = 10000, l = 1, k = 4$ が最大精度となることが分かった。また、L2 正則化を適用することで過学習が緩和され精度向上したが、教師データ量不均衡に対して、教師データ量や損失関数による調整を用いても精度向上に至らなかった。状況に応じて適切な手法を用いることが重要である。

今後は更なる精度向上を図ると共に、視聴者の年代推定だけでなく性別や世帯推定に応用していくことを検討する。

### 参考文献


[1] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting.

[2] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang, On Large-Batch Training For Deep Learning: Generalization Gap And Sharp Minima.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote : Synthetic minority over-sampling technique. JAIR, 16(1): 321-357, 2002

[4] I. Mani, I. Zhang. kNN approach to unbalanced data distributions:a case study involving information extraction. In Proceedings of wookshop on learning from imbalanced datasets, 2003.

[5] Xiaojin Zhu. Semi-Supervised Learning Literature Survey. Computer Sciences TR 1530, 2008