4ZD-01

アクティブフィンガープリンティングを用いた多値分類による会員ＩＤの推定

山本　知輝 †　　礒貝　竜真 ‡　　渡名喜　瑞稀 ‡　　利光　能直 ‡　　高山　眞樹 ‡
齋藤　孝道 †
明治大学 †　　明治大学大学院 ‡

1　はじめに

ブラウザフィンガープリンティング（以後，フィンガープリンティングと呼ぶ）とは，ブラウザから取得可能な情報（以後，特徴点と呼ぶ）を収集し，特徴点を複数組み合わせたものの差異をもとに端末を識別する技術である．フィンガープリンティングの応用として，不正検知を目的とし，会員制 Web サイトにおける会員 ID の推定技術が提案されてきた．本論文では，アクティブフィンガープリンティングを用いた多値分類による会員 ID 推定手法を提案し，5,885 件のアクセスデータを用いた検証を行った．本論文の会員 ID 推定の概念図を図 1 に示す．

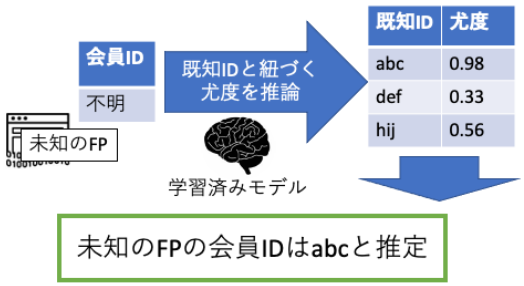

図 1: 会員 ID 推定の概念図

2　アクティブフィンガープリンティング

フィンガープリンティングは特徴点の収集方法により，JavaScript 等を利用するアクティブフィンガープリンティングと，HTTP ヘッダを用いて採取するパッシブフィンガープリンティングの 2 つに分類できる．UserAgent（以後，UA と呼ぶ）はパッシブフィンガープリンティングにおいて重要度の高い特徴点の一つであるが，Chrome において UA が段階的に固定されていくことが予見される．それに伴い，UA に依存したフィンガープリンティングは今後難しくなり得る．

フィンガープリンティングによる ID 推定の研究事例として，Park ら [1] は，アクティブフィンガープリンティング及びパッシブフィンガープリンティングにより採取できる特徴点を用いて会員 ID 推定を行った．推定の精度は，4 節で定義する会員 ID 値推定の精度で示すと，Chrome からのアクセスデータのみを用いたモデルで 89%，Internet Explorer からのアクセスデータを用いたモデルで 96%であった．

田邊ら [2] は端末の識別に有効な特徴点の組み合わせを調査した．結果として UA を含む特徴点の組み合わせを用いて最良の識別精度を達成しており，精度は AUC を用いて 0.9897 であった．

3　実験方法

3.1　データセット

本論文では，2021 年に収集した 5,885 件のアクセスデータを用いた．このデータセットに含まれる会員 ID の数は 1,582 であった．本論文では収集サイト初回アクセス時に発行している Cookie を会員 ID として扱う．

本論文では，アクティブフィンガープリンティングにより収集できる特徴点と，これらをもとに生成した特徴点を加えた総数 71 の特徴点を利用する．なお，今回の実験ではパッシブフィンガープリンティングにおいて会員 ID 推定に重要な特徴点である IP アドレス，UA を使用せずに会員 ID 推定を行なった．

アクセスデータのうち 8 割を学習用，2 割をテスト用とした．以後，学習用データ内で出現した会員 ID を既知 ID，それ以外の会員 ID を未知 ID とする．

3.2　モデル作成

LightGBM を用いた多値分類により，テスト用データセット内のアクセスデータの会員 ID が，各既知 ID と一致する尤度を求めるモデルを作成した．学習用データの 8 割をモデルの学習に用い 2 割を検証に用いた．精度の評価にはテスト用データを用いた．学習用データおよびテスト用データの件数と会員 ID 数を表 1 に示す．なお，テスト用データ内の会員 ID のうち 433 件は学習用データ内の会員 ID と重複している．

表 1: 学習用データ及びテスト用データのデータ数と会員 ID 数

	データ数	会員 ID 数
学習用データ	4,708	1,397
テスト用データ	1,177	618

ID Estimation Using Active Fingerprinting by Multi-class Classification
†Tomoki YAMAMOTO　‡Tatsuma Isogai　‡Mizuki TONAKI
‡Yoshinao TOSHIMITSU ‡Masaki TAKAYAMA †Takamichi SAITO
†Meiji University
‡Graduate School of Meiji University

会員ID1件当たりのアクセスデータの件数の平均値，最大値，最小値を表2に示す．

表2: 会員ID一件あたりのFPの件数

	平均値	最大値	最小値
全アクセスデータ	3.71件	48件	1件
学習用データ	3.37件	37件	1件
テスト用データ	1.90件	11件	1件

3.3 会員ID推定

「会員ID推定」は以下の手順で行う．まず，未知のアクセスデータと紐づく既知IDが存在しないかを推定する（以後,「会員IDなし推定」と呼ぶ）．紐づく既知IDが存在するとした場合，紐づく会員IDの値を推定する（以後,「会員ID値推定」と呼ぶ）.「会員ID値推定」と「会員IDなし推定」を併せたものを「会員ID推定」とする。

以下に具体的な手順を示す．

1. テスト用データ内の各アクセスデータを，3.2節で作成したモデルで推論し，各既知IDと一致する尤度を求める
2. 最も大きい尤度の値が閾値以下である場合，会員IDなしと推定し会員ID推定を終了する
3. 最も大きい尤度の値が閾値より大きい場合，その会員IDを，会員ID値推定の推定結果とする

4 実験結果及びその考察

4.1 評価指標

会員ID値推定，会員ID無し推定および会員ID推定の精度を求める式を以下に示す．ただし，会員ID値推定正は会員ID値推定の正解数，会員ID値推定誤は会員ID値推定の不正解数，会員ID無し推定正は会員ID無し推定におけるTP，会員ID無し推定誤は会員ID無し推定におけるFPである．

なお，会員ID無し推定におけるTPとは会員IDが存在しないと推定し，その推定が正しかった場合を示し，FPとは会員IDが存在しないと推定し，その推定が誤っていた場合を示す．

会員ID値推定の精度 =

$$\frac{\text{会員}\mathit{ID}\text{値推定正}}{\text{会員}\mathit{ID}\text{値推定正} + \text{会員}\mathit{ID}\text{値推定誤}}$$

会員IDなし推定の精度 =

$$\frac{\text{会員}\mathit{ID}\text{なし推定正}}{\text{会員}\mathit{ID}\text{なし推定正} + \text{会員}\mathit{ID}\text{なし推定誤}}$$

会員ID推定の精度 =

$$\frac{\text{会員}\mathit{ID}\text{無し推定正} + \text{会員}\mathit{ID}\text{値推定正}}{\text{推定したアクセスデータ数}}$$

4.2 推定結果

実験の結果，会員ID値推定（433件），会員ID無し推定（181件）および会員ID推定の精度は表3のようになった．

表3: 会員ID推定の精度 (少数第6位を四捨五入)

	会員ID値推定	会員IDなし推定	会員ID推定
精度	0.93672	0.68836	0.87511
推定数	946	231	1,177

表3から，パッシブフィンガープリンティングにおいて重要な特徴点を使用せずに会員ID推定を高精度で出来たことがわかる．

4.3 考察

多値分類モデルを利用し，会員ID推定において重要な特徴点の調査を行った．その結果，特徴点の重要度について偏りが小さいことがわかった．そのため，特定の特徴点の過剰な影響による会員ID値の誤推定が減少し，高精度で会員ID推定を行うことができたと考えられる．

重要度が低い情報を削除し，会員ID推定に最適な特徴点の組み合わせを選択することで，さらなる精度の向上となるかの試みは今後の課題としたい．

5 まとめ

本論文では，5,885件のアクセスデータを用いて，多値分類を利用した会員ID推定を行った．結果として，1,177件のアクセスデータの会員ID推定を約87%の精度で行うことができた．

謝辞

本研究の成果の一部は，JSPS科研費18K11305の助成を受けたものです．また，本研究はレンジフォース株式会社の支援により実施しています．

参考文献

[1] Park Sohee, Jang Jinhyeok and Choi Daeseon (2020), ”A Study on User Authentication Model Using Device Fingerprint Based on Web Standard”, Journal of the Korea Institute of Information Security & Cryptology, Volume 30, Issue 4, Pages.631-646

[2] 田邉 一寿，高橋 和司，安田 昂樹，種岡 優幸，細谷 竜平，小芝 力太，齋藤 祐太，齋藤 孝道，2017 Browser Fingerprintingにおける特徴の組み合わせに関する考察, コンピュータセキュリティシンポジウム（2017）