課題説明ドラフト: パケットフローからのアプリケーション推定

課題の要点

主タスク	主評価指標	リーク対策	2タスク
Flow classification 暗号化フローからアプリ種別を推定	**Macro F1** 少数クラスも重視	**No DNS/SNI/IP** 直接的な手掛かりを特徴量から除外	**Static + realtime** 各50点

公開課題情報

- 課題Webサイト: https://competition.aiforgood.itu.int/web/challenges/challenge-page/495/overview
- 参加者向け公開リポジトリ: https://github.com/FG-AI4H/application-inference-packet-flows-public
- 公開日: 2026年8月1日
- 応募締切: 2026年10月23日
- インセンティブ: 賞金 1,000 USD、受賞者証明書、およびTTCからの副賞を予定しています。

背景

モバイルアプリケーションの急速な多様化により、アプリケーションごとに求められる帯域、遅延、信頼性は大きく異なるようになっています。動画配信、SNS、ブラウジング、メッセージング、地図、通知処理などは、それぞれ異なる通信パターンを持ちます。複数のアプリケーションの通信が同一のネットワーク経路を共有する場合、アプリケーション種別を考慮しない制御は輻輳、遅延、パケットロスを引き起こし、ユーザの Quality of Experience (QoE) を低下させる可能性があります。

ネットワーク運用者がアプリケーション種別を早期に把握できれば、アプリケーションに応じたQoS制御、トラフィック優先制御、ネットワークスライシング、帯域割当、異常検知などをより適切に行うことができます。一方で、現在のモバイル通信の多くは TLS や QUIC によって暗号化されており、通信内容を直接参照する従来型の識別は困難です。そのため、ペイロードに依存せず、パケット長、到着間隔、通信方向、フローの初期挙動などの統計的特徴からアプリケーションを推定する技術が重要になります。

さらに、モバイルネットワークではリアルタイム性と計算資源の制約も重要です。通信開始直後の短時間に多くのパケットが集中するアプリケーションでは、識別が遅れるとQoS制御の効果が低下します。一方で、すべてのパケットに対して重い特徴抽出や推論を行うことは、処理遅延やエネルギー消費を増加させ、実運用上の負担となります。したがって、高精度な識別だけでなく、少ない情報から早期に推定する能力、誤分類を抑制する設計、計算資源を効率的に使うモデル設計が求められます。

本課題の目的

本課題の目的は、暗号化されたモバイル通信から抽出したフロー特徴量を用いて、各通信フローがどのアプリケーションに由来するかを推定するAI/MLモデルを構築することです。参加者は、通信内容やドメイン名に直接依存せず、パケット系列の統計的特徴からアプリケーションを識別する手法を設計します。

参加者は、提供された学習用データを用いてモデルを設計し、評価用データに含まれる各通信フローのアプリケーションラベルを予測します。DNS名、TLS SNI、HTTP Host、IPアドレスなど、ラベルに直接結びつく情報は特徴量から除外されています。したがって、参加者は通信量、通信方向、パケットサイズ、到着間隔、バースト性、フロー継続時間などを活用して推定を行う必要があります。

本課題では、単に分類精度を高めるだけでなく、実ネットワークへの適用を意識した手法を期待します。例えば、通信開始直後の少数パケットからの早期識別、確信度が低い場合の判断保留、計算負荷の小さいモデル、クラス不均衡に強い学習方法などが重要な観点になります。これにより、リアルタイムQoS制御やアプリケーション単位のネットワーク管理に利用可能な、資源効率の高いアプリケーション識別技術の実現を目指します。

アプローチ

参加者は以下を自由に設計できます。

- 使用する特徴量の選択
- 特徴量変換、正規化、次元削減
- AI/MLモデルの選択
- クラス不均衡への対処
- 少数サンプルのアプリケーションに対する汎化手法

想定されるアプローチには以下があります。

1. フロー統計ベースのアプローチ

パケット数、バイト数、通信方向、フロー継続時間、パケットサイズ分布、到着間隔などの統計量を用いて分類モデルを構築します。

2. 時系列ベースのアプローチ

パケット列を時系列データとして扱い、パケットサイズ列、方向列、到着間隔列などを用いてモデル化します。

3. ハイブリッドアプローチ

フロー統計量と時系列特徴量を組み合わせる、またはドメイン知識に基づく特徴量設計と機械学習モデルを組み合わせる方法です。

タスク

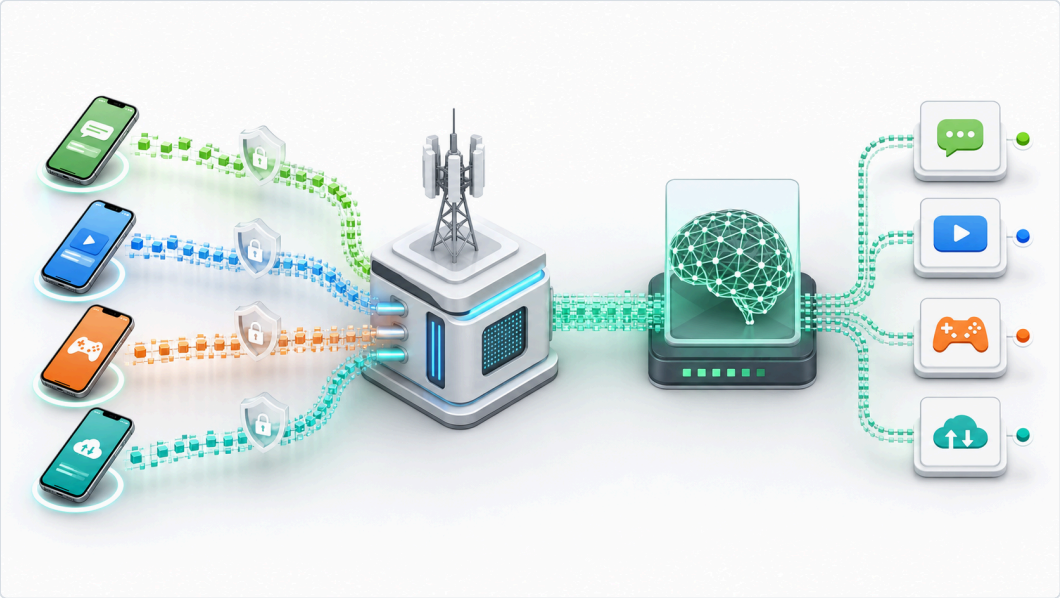


図1: 暗号化パケットフローからアプリケーション種別を推定する課題の概念図

本課題では、同じアプリケーションラベルを対象とする2つのタスクを設定します。タスク1はフロー全体を用いる静的推定、タスク2は通信開始直後の情報だけを用いるリアルタイム推定です。

タスク1: 静的アプリケーション分類

評価データに含まれる各フローについて、フロー全体から計算した統計特徴量を用いて、学習データに含まれる既知アプリケーションのいずれかを予測します。

使用するデータセット:

- `dataset/task1_static/`

タスク2: リアルタイムアプリケーション分類

評価データに含まれる各フローについて、通信開始直後の最初の5パケットから計算した特徴量だけを用いて、既知アプリケーションのいずれかを予測します。

使用するデータセット:

- `dataset/task2_realtime_prefix5/`

タスク2では、タスク1のフロー全体特徴量を使って予測を作成してはいけません。タスク2の予測には、`task2_realtime_prefix5/` に含まれるファイルのみを使用してください。

主な対象ラベル:

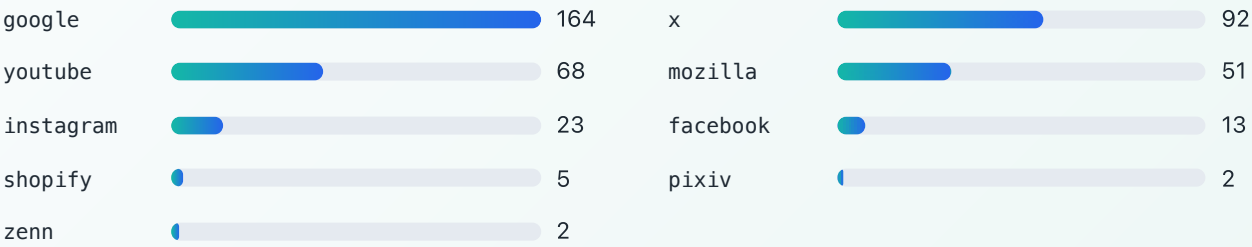
- `facebook`
- `google`
- `instagram`
- `mozilla`
- `x`
- `youtube`

データセット

データセット構成

Task	用途	Train	Validation	Test	点数
`task1_static`	静的推定: フロー全体特徴量	247	83	81	50
`task2_realtime_prefix5`	リアルタイム推定: 最初の5パケット特徴量	247	83	81	50

閉集合データセットのラベル分布



元データは、単一UEの通信を約44分間キャプチャした pcapng ファイルです。pcapng からフロー単位の特徴量を抽出し、CSV形式で参加者に提供します。

各タスクディレクトリには以下のCSVが含まれます。

- `train_features.csv`: 学習用フロー特徴量
- `train_labels.csv`: 学習用フローの正解ラベル
- `validation_features.csv`: 検証用フロー特徴量
- `validation_labels.csv`: 検証用フローの正解ラベル

- `test_features.csv`: 評価用フロー特徴量
- `sample_submission.csv`: 提出ファイルのテンプレート

主催者のみが保持するファイル:

- 最終評価用フローの正解ラベルは主催者のみが保持し、参加者には配布しません。

特徴量には、以下のような情報が含まれます。

- フロー継続時間
- トランスポートプロトコル
- サーバ側ポート
- 上り/下りのパケット数
- 上り/下りのバイト数
- パケットサイズの平均・標準偏差
- パケット到着間隔の平均・標準偏差・最大・最小
- バースト数
- 上り/下り比率

以下の情報は参加者配布用特徴量には含まれません。

- IPアドレス
- DNSクエリ名
- TLS SNI
- HTTP Host
- URLまたはドメイン文字列
- ペイロード

評価方法

> **評価の中心:** クラス不均衡の影響を抑えるため、主指標はmacro F1、補助指標はaccuracyとします。

参加者は、2つのタスクそれぞれについて、評価用フローごとのアプリケーションラベルを予測して提出します。課題プラットフォーム上での推奨提出形式は、`task1_static` と `task2_realtime_prefix5` の2つのトップレベルオブジェクトを持つJSONファイルです。

推奨JSON提出形式:

```
{
  "task1_static": {
    "flow_000001": "youtube",
    "flow_000002": "instagram"
  },
  "task2_realtime_prefix5": {
    "flow_000001": "youtube",
    "flow_000002": "instagram"
  }
}
```

評価スクリプトは、以下2つのCSVファイルを含むZIPファイルも受け付けます。

- `task1_static.csv`
- `task2_realtime_prefix5.csv`

各CSVファイルには、対応するタスクの予測ラベルを記載します。

```
flow_id,app
flow_000001,youtube
flow_000002,instagram
```

主評価指標は macro F1 とします。クラス数およびサンプル数に偏りがあるため、単純な accuracy だけでなく、少数クラスを含めた分類性能を評価します。

参考指標として accuracy も算出します。

公式スコアは100点満点とし、タスク1とタスク2をそれぞれ50点として評価します。各タスクの点数は macro F1 に基づいて算出します。

- タスク1: 50点
- タスク2: 50点

参加者は公開されている `validation_labels.csv` を用いて、自分の手法をローカルに検証できます。最終評価用の正解ラベルは主催者のみが保持し、参加者には配布しません。

条件

- 評価用データの正解ラベルを学習に使用してはいけません。
- 最終評価用の正解ラベルは参加者に配布してはいけません。
- タスク2の予測にタスク1のフロー全体特徴量を使用してはいけません。
- 配布された特徴量以外の情報を利用する場合は、提出資料に明記してください。
- pcapng を追加配布する場合、DNS名、TLS SNI、HTTP Host などから直接ラベルを推定する方法は、主タスクの趣旨とは異なるため、別トラックとして扱うことが望ましいです。

提出物

参加者は以下を提出します。

1. 予測ファイル

評価用フローに対する予測ラベルを含むJSONファイル、またはタスクごとのCSVファイルを含むZIPファイル。

2. ソースコード

学習、推論、提出CSV生成に用いたコード。主催者は、非公開ラベルや禁止された特徴量を使用していないか確認します。

3. 説明資料

PDF形式で、以下を含めます。

- 使用した特徴量
- モデル構成
- 学習方法
- クラス不均衡への対処
- 評価結果
- 提案手法の新規性
- 計算コスト

評価基準

評価基準 #1: 推定精度

macro F1 を主指標として評価します。accuracy も参考値として確認します。

評価基準 #2: 汎化性能

少数クラスや通信量の少ないフローに対しても安定した推定ができる手法を高く評価します。

評価基準 #3: モデル設計

特徴量設計、モデル構成、クラス不均衡への対処、過学習抑制などを総合的に評価します。

評価基準 #4: 実用性

計算負荷が小さく、実ネットワーク運用に適用しやすい手法を高く評価します。