MWS Cup 2021 課題3 解説

株式会社ＦＦＲＩセキュリティ
（東証マザーズ：3692）
https://www.ffri.jp/

2021-10-26

作問メンバー

株式会社FFRI セキュリティ

押場 博光

茂木 裕貴

中川 恒

＋ デバッグプレー協力者

アジェンダ

1. 問題構成
2. 解説・意図
3. その他施策
4. 反省

問題構成

1. 表層解析ログ分析

- **課題データを事前に分析しておくことを推奨**

2. 表層解析ログ分類

マルウェア/クリーンウェア分類

- 課題データを対象に**分類器を事前に検討しておくことを推奨**
- **新たな試みとして1週間前からSubmitを可能に**
- 良い成果は将来的な論文化やGitHubでの公開を期待

事前公開している課題データ(FFRI Dataset 2021と同形式)を使用

問題構成

表層解析ログについて

検体を動かさなくても入手可能

- 解析コストが比較的低い場合が多く大量の検体の解析が可能
- 実行時のコンテキストを意識しなくてよい
  - データセットの拡張・分析がしやすく非常に使いやすい
- マルウェアを実行前に検知し得る

但し，**情報量は劣る**

最大限活用してもらいたい

問題構成

表層解析ログの分析問題

表層情報から特徴量を検討するための一助

- **後半の分類問題に特徴量を活かしてもらう意図**
- 主にドメイン知識をきちんと分析する話をちゃんとやってもらいたい
- ツールの出力を鵜呑みにせず必要な処理を行う

表層解析ログの分類問題

実際的なマルウェア検知モデル構築の体験をしてもらう

問題構成

課題3の配点

種別	課題	配点
分析	ファイルの収集日	1
	初期化データセクションの合計サイズ	1
	文字列の最大長	1
	性質X1, X2, X3	1
	性質S2を満たす割合	1
	パック判定とセクション	1
	パック判定と.NET	1
	パック判定とセクションエントロピー	1
分類	分類問題	12

計 20点

想定回答について

同時に想定解法を配布しますので、基本的にはそちら参照頂ければと思います！

本資料内では、それ以外の部分を解説します

分析問題

基本的には解答例をご覧いただければOK

全体を通じてのコンセプトは、ドメイン知識をベースに特徴量を作成すること

JSONの項目を放り込むだけでなく、ドメイン知識に基いて特徴量を作成する

必要ならツールの出力値を補正する

課題データについて

現実世界でのマルウェア検知を意識して課題データを作成・評価

現実の検知製品におけるマルウェア分類では99.9%以上がクリーンウェア

- **少ないマルウェアを検知し漏らすと製品としての存在意義がなくなる**
- **過検出は利用者にとって大きな損害となる**
  - **FPRが低かったとしてもFPの絶対数は大きくなりやすい**

現実の「評価」を意識した精度評価

- お客様はマルウェアをいくつか用意して検知できるか複数製品を試すことがある
- 過検出が多ければお客様は製品を解約する

課題データについて

test.jsonl

不均衡データ

- マルウェアが300件、クリーンウェアが2100件
- ただし評価はそれぞれ別

分類問題 コンセプト

イメージ	PB	概要	想定配点
ベースライン	0.61以下	your_first_submission.csvとほぼ同等かそれ以下	1
…	0.61以上 0.95未満	分析問題の特徴量を活用できた	2~8点
頑張ったとこ	0.95以上	データ分析がんばったチーム	9

+ 順位点予定

その他施策

背景

前々回、取り組んでもらえないチームがいくつか存在した

- **敷居が高すぎるという声もあり何か施策が必要と認識**

施策

1. データの事前公開
2. ベースライン実装のご提供
3. 分析環境のご案内
4. 周辺ツールのご提供
5. **データの削減（今回から）**
6. **1週間前からSubmit可能に（今回から）**

その他施策 - データ事前公開

前回からデータを事前に公開し予習・試作を可能に

前々回はデータをパスワード付zipにて事前提供し、当日パスワードをご案内

- **予習するにも巨大なFFRI Dataset 2019を利用する他なく事前準備が難しい**
- **結果的に当日大きなデータを分析することとなりかなりハードルが高かったと反省**

- 当日は評価結果をもとに精度を向上していくオペレーションを想定

今回の反省としては、もうちょっと早く実施できればよかったという点

- **事前案内 / データ公開**
- **1週間前からSubmit可能なため、さらに早いほうがよかった**

その他施策 – ベースライン実装のご提供

サンプルのJupyter Notebook(+CSV)をご提供した

課題

- 前々回、登録形式に関するご案内に問題があった
- また、この手の課題に取り組んだことのない人向けの配慮が必要と反省

取り組んでいただくための第一歩になることを期待

- データと共に事前提供することで当日行うべき工程のイメージを持ってもらえるように
- 今回取り組んでいただけなかったとしても、次回以降の参考となるように
  - まずはこれさえご投稿いただければ、1点は付与する想定だった

その他施策 – 分析環境

分析環境としてGoogle Colaboratoryを許可

ローカルのPC/サーバーで分析環境を用意するのは割とハードルが高い認識

- この点で参加できるチームに制約がかかるのは本意ではない

そこで、皆一定水準以上の環境で分析できるようにしたいと考え検討

- 無料で利用できるGoogle Colaboratoryの利用を許可
- これを実現するため、学習データ量を削減(後述)
- (他の環境を利用したい場合には事前にご相談頂くようご案内)

その他施策 – 周辺ツールのご提供

前処理をお手伝いするPython package “FEXRD”のご提供

機械学習等のための前処理についてはある程度知見が必要

- 我々として競ってほしいのはこの点ではない
  - 機械学習特有の部分ではなく、マルウェア分類ならではの部分で競ってもらいたい

その他施策 – データの削減

学習データの量を削減した

学習データが多いとGoogle Colaboratoryで分析が厳しい

- 前回はDaskの利用を推奨したが、Daskに慣れているかどうかで結果が決まってしまうという声が出た
  - 我々として競ってほしいのはこの点ではない(FEXRDの項と同じ)
- Pandasで処理できる量にデータ量を削減

データ量を削減し続けることはできないので、この点は次回以降の課題

その他施策 – 1週間前からSubmit可能に

競技当日の1週間前からSubmit可能に

Kaggleなどの機械学習コンペティションの多くは一定の期間開かれている

- スコアボードが別なため、本課題独自の取り組みとして導入
- より取り組みやすいよう、全体のSubmit上限のみを設けた
  - 1日毎のリセットではないため、特定の日に集中して取り組んでも良い

反省

問題形式

問題形式は昨年と同様

一方で、問えるフィールドにも限りがあり、より面白い問題にしていくのが難しい

- 次回以降は分類問題だけにするなど、形式を変更することも検討

課題データ・スコアボード公開時期

1週間前から提出できるようにしたことは悪くないと思っている

一方で、課題データとスコアボードの公開はそこまで早くなく、登録期間が短かった

- 次回はもっと早く公開できればと思う