



MWS Cup 2024 マルウェア分類

株式会社 F F R I セキュリティ
<https://www.ffri.jp>

作問メンバーについて

紹介

今年の作問体制



今年は以下のメンバーにご協力いただきました（敬称略）

所属・氏名		役割
株式会社 FFRI セキュリティ	茂木裕貴	作問チームリーダー
株式会社 日立システムズ	石田裕貴	作問委員
早稲田大学	杉山孔亮	作問委員
	池澤隆人	作問委員

今年の問題について

背景

今年の問題の方針



今年の問題はマルウェアの動作の予測とした

	2023	2024
データの 種類	FFRI Datasetの 形式ほぼ全項目＋ マルウェアの ファミリラベル	FFRI Datasetの 形式ほぼ全項目＋ MalwareBazaarの 解析API
問題	マルウェアの ファミリ分類	マルウェアの 動作の予測

表層情報から動作を推定

推定したいシグネチャ以外のシグネチャも提供

特徴

**表層情報による
動作の推定**

**推定したいもの
以外のシグネチャ**

概要

「動かした」結果を
「動かさずに」知れたら良い

マルチラベル推定などを用いて
他のラベル情報を活用できるか

ラベルに関する注意

今回動作ラベルはMalware BazaarのAPI（のうちvxCube）をそのまま使用した

あくまでサンドボックス上の解析の結果のため、「本当にある動作が行われない」のか「本来はある動作を行うがたまたまサンドボックス上では発現しなかった」のかは分からない

後者の場合、訓練したモデルは表層情報から「この動作は行われない」と予測してしまうが、実際にはその動作があり得る

理想はハンドラベリングだが、それはかなりの高コストである

とはいえサンドボックスで動かすのもそれなりに手間であるため、その手間を表層情報からの予測で省けるのであれば意義はある

例えば昨年のファミリラベルはマルウェアを動かさずにyaraで取ったものなので、その点でいうと今回の方がタスクとしての意義はある（と思う）

採点とヒントの方針

例年の二極化を改善しデータコンペの初心者にも取り組みやすくするため、配布するNotebookのヒントを詳細に書き、ヒントを全部使うと満点（Writeup点及び順位点除く）がとれるようにした

```
# Second Submission
# 特徴量の欠損値を0で埋めてみる
# それなりにスコアが向上する
X_train = # ここを埋める
cv(X_train, y_train, lambda: Rand
X_test = # ここを埋める
df_pred = make_prediction(X_train
df_pred.to_csv("second_sub.csv",
```

```
# Fourth Submission
# ClassifierChainを使う
# ClassifierChainは複数のラベルを予測する際、まずあるラベルxを予測し、
# 次に予測したxの値を特徴量に追加してラベルyを予測し、次にxとyの予測値を
# 特徴量に追加してzを予測し、とやる手法
# ラベル間に相関があれば単にそれぞれのラベルを予測するより精度が向上しうる
# ただし、xの予測精度が悪いと、そのエラーがyやzの予測にも影響しうる
# そのため予測の順番も重要になってくる
# 順番を決めるためまずは各シグネチャにおけるcvの値を見る
```


問題の回答について

結果

データを見ることの重要性

今回模範解答は配布したNotebookの穴埋めに留めている

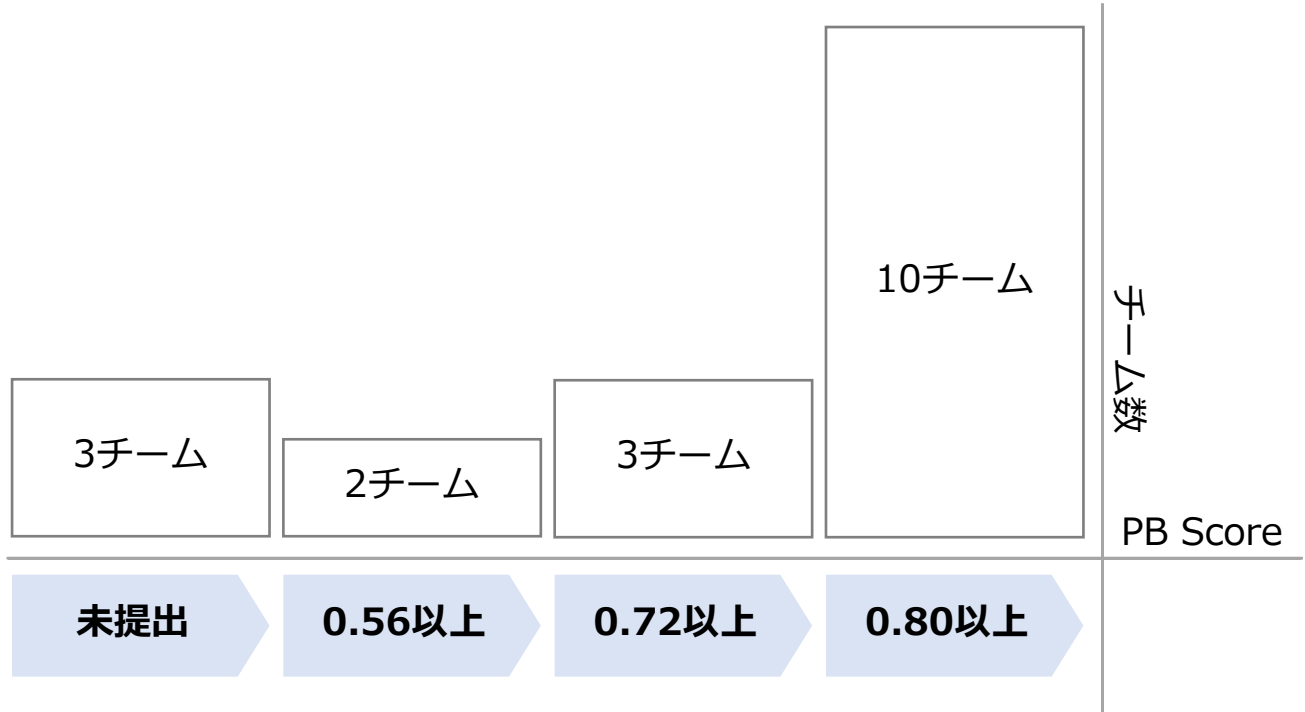
そこまでできれば20点（Writeup点と順位点を除いて満点）

データを見るとラベルに対しかなりの不均衡データになっていることが分かる

PB 0.80を超えたチームはどこもここに気づいていた感触

配布したNotebook以上の点がほとんど

取り組んだチームのほぼすべてが20点(0.72以上)は越えた



※配布したNotebookを穴埋めするとPBは0.727程度だが、多少変動があるため、0.72以上を20点とした¹¹

ヒントは機能していた

今回はかなりヒントを詳しく書いた

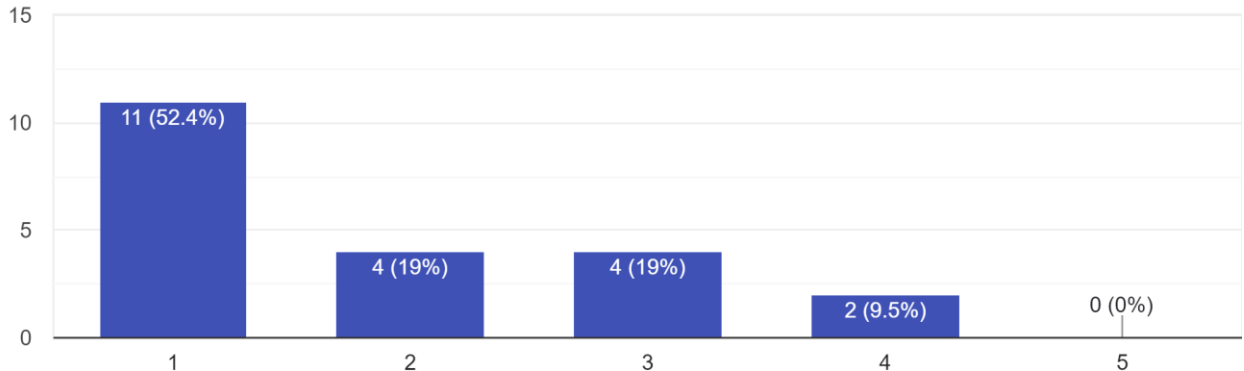
ほぼ穴埋めで、穴埋めも（最後以外は）難しくない

充実しすぎというフィードバックもあった

（1が参考になった、5が参考にならなかった）

ヒントは参考になりましたか

21件の回答



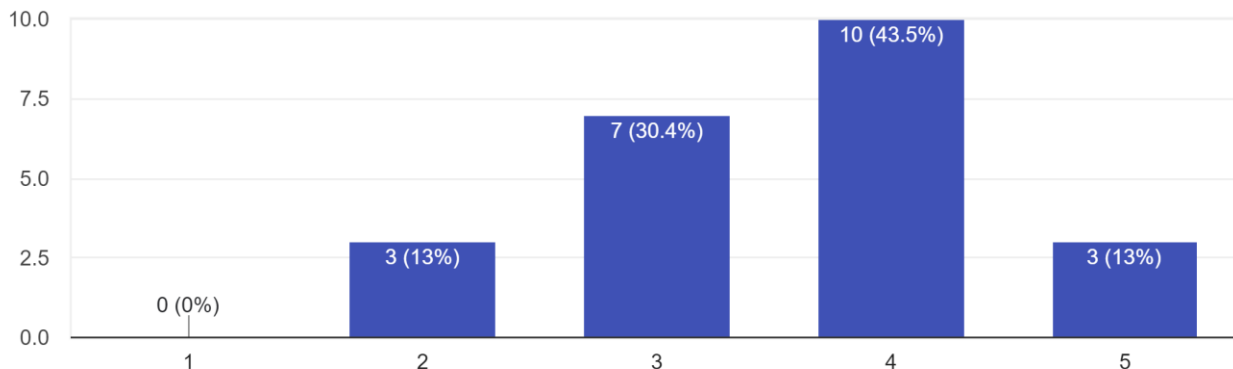
例年より簡単の方向に倒れた

ヒントを詳細に記載した点、およびタスクとしては2値分類問題であることも要因に思われる

(1が簡単、5が難しい)

表層解析課題の難易度はどうでしたか？

23件の回答



Writeupはクオリティが高かった

グラフを使って可視化するなど、データをきちんと観察できているチームが多かった

Writeup自体もとても良く書けているチームが多かった

フリーコメント返し

FFRIデータセットやFEXRDに関してもう少し説明が欲しい

- FFRIデータセットの説明文は付け足す

- FEXRDはリンク先にドキュメントとコードがあるのでそれを参考にしてほしい

Writeup強制公開

- 参加ハードルが上がりそう

- 任意ですが、皆さま是非公開の方をよろしくお願いいたします

中間Notebook発表

- 難しそう（良いものを伏せるなどHackできる）

順位のポイント変動が少ない

- 順位点はどう頑張っても競技の相手次第で取れない点数のため、割合を増やすのが正しいかはなんとも

- ここまでヒントが詳細であれば順位点を増やしても良いかも

Towards : 次回以降



作問にご協力いただける方を募集しています！！！！

