MWS Cup 2018 課題 3 振り返り

2018/12/20

アジェンダ

- Q1. 動的解析ログによるコードインジェクション分析
- Q2. 表層解析ログによるファイル分類
- Q3. データセット改善提案

Q1. 動的解析ログによるコードインジェクション分析

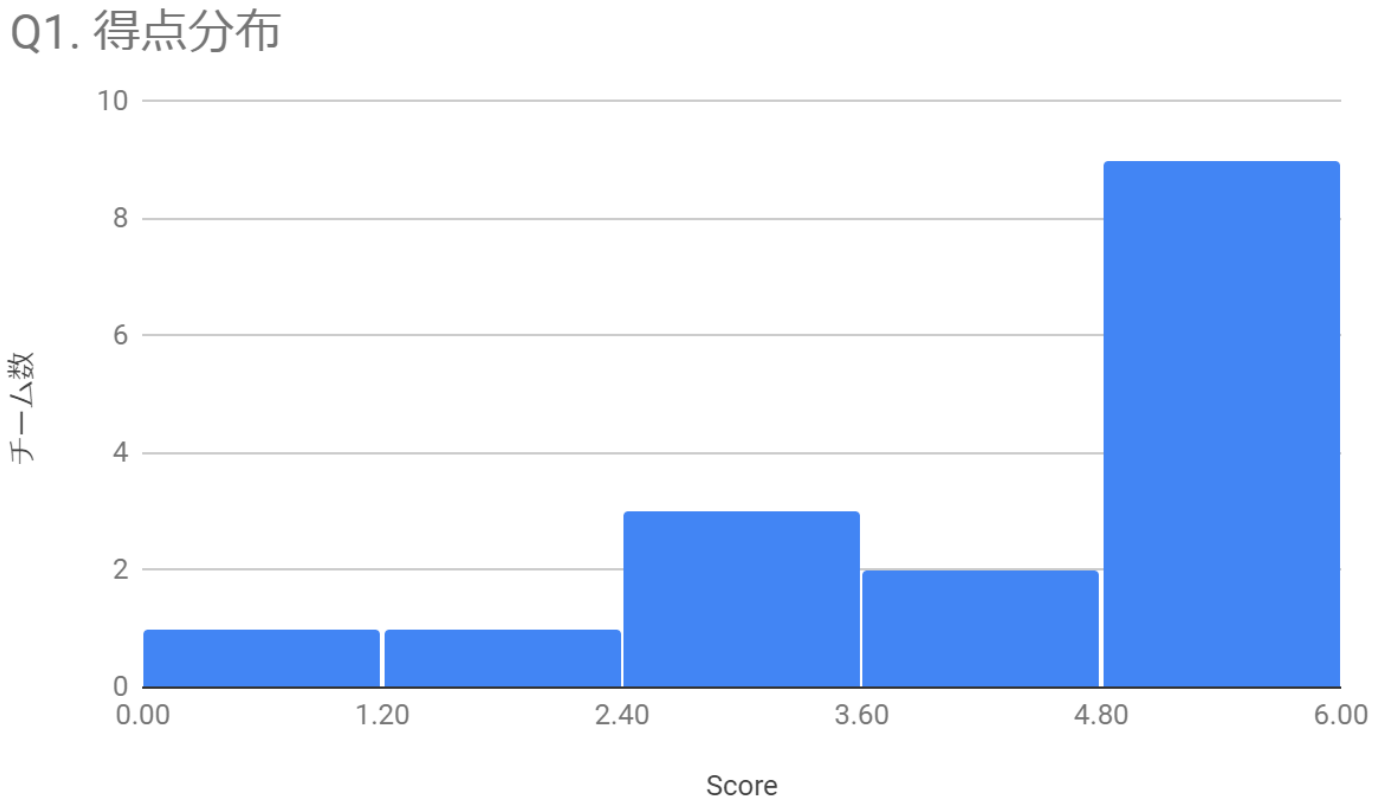
- **出題内容**
  - Cuckoo Sandbox のログである 3 つの json ファイルを分析し、マルウェアによるコードインジェクションの詳細を明らかにする
- **出題意図**
  - 過去問の解法解説を見て勉強し、ログを分析するちょっとしたスクリプトを書いて欲しい

> **基本的にはコードインジェクションに使われる API と呼び出し順序を理解し、引数が process_handle != 0xFFFFFFFF の API 呼び出しを追っていくことで解ける**

- **過去問との違い**
  - API 名だけでなく、スレッドやセクションのハンドル等の引数やインジェクション手法の名称を問う
  - 約 8,000 件のログファイルの中から指定したログと最も似たパターンのログを探す

Q1. 結果

- 満点のチームはなかったが、全体的に得点率は高く、過去問対策していれば簡単だったと思われる



Q1. 良かった点・悪かった点

- **良かった点**
  - 事前課題に取り組むきっかけとなったこと
    - 事前課題の成果物として、初心者でも Cuckoo のログ分析ができる Web UI やフィルタリング機能を持つ実用的なツールが開発された
- **悪かった点**
  - 過去 3 年間、同じ種類のコードインジェクションを分析対象にし続けているため、理解している人にとっては新たな学びが少ない
- **今後**
  - 同じ形式なら、これまで扱っていないコードインジェクション手法を分析を題材にする
  - 動的解析のログ分析ではなく、実際に手を動かして動的解析したり、Cuckoo のシグネチャを作ったりする課題はどうか？

Q2. 表層解析ログによるファイル分類

- **出題内容**
  - FFRI Dataset 2018 と同形式のファイル表層解析データ 10 万件をクリーンウェアデータとマルウェアデータに分類する
- **出題意図**
  - 仕様変更した FFRI Dataset 2018 に触ってもらいたい
  - FFRI Dataset 2018 の 50 万件のデータを事前に分析し、分類手法の検討や分類器の作成をしてほしい
  - マルウェア対策技術の開発では誤検出に対する要求が厳しく、誤検出率を下げる工夫が重要であることを知ってほしい
- **得点計算式**
  - 10 万件のデータ分類結果の TPR(検出率)と FPR(誤検出率) に基づいて、12 * (TPR - 50 * FPR)
  - 誤検出率が高いと大幅減点

Q2. 手法解説の採点基準

- 分類精度以外に、手法を解説する資料を提出してもらい評価

基準	得点
無回答	0
とにかくやってみました	1
データ分析をしっかり行って理論的にロジックの検討を行っている	2
事前の精度検証や FPR 低減を意識した工夫や試行錯誤がある	3

Q2. 結果

- 今回この課題で点を得たのは 16 チームのうち、「ICSCoE2018」 と 「Team GOTO Love」 の 2 チームのみ
- 両チームとも FPR は 0.1 % 台、得点差は僅か 0.001 のほぼ同点
- 正解ラベルは、クリーンウェア 50,000件、マルウェア 50,000件

Team	Score	TPR	FPR	TP	FP
ICSCoE2018	8.296	78.636%	0.19%	39,318	95
Team GOTO Love	8.295	75.728%	0.132%	37,864	66

- 多くのチームが機械学習を用いて分類に挑戦していたが、FPR が高すぎた
- マルウェアのみの特徴だけでなく、クリーンウェアの特徴にも目を向け、機械学習のみに頼らず、パターンマッチングなどと組み合わせることで FPR を下げられた可能性がありそう

Q2. 良かった点・悪かった点

- **良かった点**
  - 仕様変更した新しい FFRI Dataset 2018 をよく分析してもらえた
    - 全チーム無得点の可能性も考えていた
  - スコアサーバーを用意し、競技中に全チームの解答状況を可視化して楽しめた（自分が）
  - 課題の取り組みが研究論文に発展

- **悪かった点**
  - 事前準備を期待していることやルールのアナウンスが不足
  - 得点計算式が厳しかったため、ぶっつけ本番のチームが戦意喪失
  - 解答状況の可視化を活かして、試行錯誤を促すヒントを提供することができなかった

Q2. 今後やりたいこと

- 試してみたいアイデア
  - ルール・得点計算式など
    - より得点するチームが増えるように順位に応じて得点を加算
  - 事前準備・ぶっつけ本番支援
    - 今回良い結果を残した手法やサンプル分類器をヒントとして共有

- 流行するマルウェアは変化していくため、
  今回の手法が今後どの程度の判定性能を維持できるのか気になる
- この課題で考案された手法が最新のマルウェアをどの程度検出できるか継続的に計測、可視化してみたい
- MWS 発の複数の手法を組み合わせて高性能なオープンソースのマルウェア検出エンジンが実現できたら面白い

Q3. データセット改善提案

- **出題内容**
  - FFRI Dataset に対して改善を提案
- **出題意図**
  - より多くの人に FFRI Dataset を使って研究をしてもらうため
- **採点基準**

基準	得点
無回答	-2
提案理由がないもの	0
提案理由が説明不足だが、検討の余地がある案	1
提案理由の説明が十分で、次期 FFRI Dataset の仕様として採用を検討したい案	4

Q3. 採用候補提案

Windows 以外の OS や携帯端末を対象とした検体の解析ログの提供

- **提案者「セキュリティ讃歌」**
- **主な提案理由**
  - 特に Android 端末を対象としたマルウェアに関するデータセットは、MWS Datasets では提供が停止されているか更新されておらず、最新のマルウェアに関する情報を入手することが難しくなっている。
- **想定されるコストやリスク**
  - 新たに解析環境を構築しなければいけない。また、Android で多く見られるアドウェアはマルウェアとの区別が曖昧であり、マルウェアかクリーンウェアかの分類が難しいという問題がある。
- **選定理由と今後の見通し**
  - コストが問題だが、表層情報の提供でもそれなりに価値がありそう

今後実現可能性を検討

Q3. 採用候補提案

FFRI Dataset を利用するための API の提供

- **提案者「SecCap-KKK」**
- **主な提案理由**
  - データセットはサイズが大きく、配布形式の変換やツールの配布では利便性の向上に限界がある。
  - API 作成によりデータを取得する際の仕様がはっきりとすれば関連ツールの開発が盛んになる。また、データセットを動的に取得することで最新のデータを取り込むことができる。
- **想定されるコストやリスク**
  - API 開発時のほか、データセットが変わった際の仕様変更のコストがかかる。
- **選定理由と今後の見通し**
  - リアルタイムなデータが欲しいという要求からAPIという提供形態を提案していることを評価

社内で提案・調整中

Q3. 採用候補提案
表層解析のデータセットへの文字列データの追加

- **提案者「Team GOTO Love」**
- **主な提案理由**
  - 実行ファイルに含まれる文字列情報からクリーンウェアとマルウェアの分類や、亜種の発見はすでに研究がなされており、実用性がある。
- **想定されるコストやリスク**
  - 実行ファイルから文字列を抽出することは strings コマンドにより短時間で行うことができるため、低コストで行うことができる。
  - 文字列情報の抽出はマルウェアを実行することなく、またインターネットに接続することなく抽出可能であるためリスクが低い。
- **選定理由と今後の見通し**
  - コストが低い点を評価

次回、表層情報を提供する場合には採用予定

Q3. 今後

- MWS では今回から Slack が常設され、データセットについて議論するチャンネル #dataset ができている
- この課題はお役御免にしたい
  - Dataset に対する提案や要望、データを取得したい検体のハッシュ値などをこのチャンネルにどんどん書き込んでもらいたい

おわりに

- 今回、新たに出題した表層データによる分類チャレンジの成果は、論文化という学術的発展やマルウェア対策製品開発のヒントとなる可能性を感じた。今後も継続してコラボレーションを考えたい。

- **WANTED**
  - MWS Cup 課題作成や企画に参加してくれる人
  - FFRI Dataset 仕様検討・作成に協力してくれる人
  - FFRI Dataset 利用の有無に関わらず、セキュリティ新技術を共同で研究開発するパートナー
  - 透明性の高い信頼されるセキュリティ技術や製品を一緒に創っていこうという善良なエンジニア

MWS の Slack で連絡待ってます！