4Z-5

サイバー攻撃検知に関わるパラメータ抽出法の検討

鈴木男人† 小池愛理† 鈴木貴之‡ 宮保憲治‡

†東京電機大学大学院　情報環境学研究科　‡東京電機大学　情報環境学部

1. はじめに

近年，社会的な問題として認識されているサイバー攻撃に対する具体的な対策が急務となっている．従来のサイバー攻撃検知手法の主流である，事前に定義したシグネチャやルールに基づいたパターンマッチング方式では，未知のマルウェアやマルウェアの爆発的な増加への対応が困難な状況である．

本稿では，サイバー攻撃による通信トラヒック（異常通信）と通常時の通信トラヒック（正常通信）パターンを，機械学習アルゴリズムを用いて解析し，効率的にサイバー攻撃検知を実現できるパラメータの評価結果を述べる．

2. 異常通信の検知手法

異常通信の検知手法として，パターンマッチング法とヒューリスティック法が挙げられる．

パターンマッチング法は，事前に通信パターンを定義し，そのパターンに該当する通信を検知する手法である．そのため，定義されていない新種の異常通信は検知できない．更に，爆発的に増加するマルウェアに対応するためには膨大な量のパターン種別の事前把握が必要となる．このため，近年のサイバー攻撃への対処手段としては，パターンマッチング法によるマルウェア検知は困難な状況である．

一方，ヒューリスティック法では通信の挙動に着目し，機械学習アルゴリズムを用いて検知を行う．新種の異常通信の検知が可能であり，事前に膨大なシグネチャを定義する必要もない．しかし，パターンマッチング法と比べて誤検知の確率が高くなる可能性があるため，新たな評価法の検討が必要となる．

3. 異常・正常通信データ解析実験

3.1. 解析データ

解析に用いるデータとして，サイバー攻撃が行われた通信データ（異常通信）と通常時の通信データ（正常通信）が必要である．異常通信の分析用に，MWS2013 で研究者用のデータセットとして CCC Dataset [1]が提供されている．データセット内には，2008 年から 2011 年までのハニーポット 2 台でキャプチャされたサイバー攻撃通信ログが含まれている．攻撃通信は，各々20 分間隔で区切られており，アプリケーションの脆弱性攻撃からマルウェア感染後の C&C サーバとの通信などが観測されている．

一方，正常通信の解析用には，研究用に提供されているデータセットを独自に収集する必要がある．収集データとしては，アプリケーション毎に異なるパラメータを含んだ通信トラヒックデータと，PC で通常活用される多様なアプリケーションの混合通信時のトラヒックデータを対象とする必要がある。

本稿では，異常通信としては CCC Dataset の 2011 年 1 月 28 日から 31 日までの 3 日間のデータを用い，正常通信としては,メール通信や動画通信などのアプリケーションごとのデータ 5 種類と，研究室の PC1 台の 4 時間分の通信データを対象とした．第一段階の評価基準として，サンプルデータ数の比率は約 1:1 の割合となるように抽出を行った．

3.2. 識別単位

正常通信と異常通信を識別する単位として，パケット単位，フロー単位，タイムスロット（TS:Time-Slot）単位での識別が可能である．

パケット単位での識別は，検知サーバに到着したパケットを遅延なく検知できる利点がある．しかし，パケット一つ一つを対象とするため識別回数が多くなる．特にバースト的に短時間に大量のパケットが到着した場合，サーバに高負荷がかかる欠点がある．

フロー単位での識別は，アプリケーションごとの一連の通信フローが終わるまで識別できないため，フロー開始から識別までに時間を要する．そのため，リアルタイム性を考慮し，各フローの先頭から数パケット分のみを用いることが効果的である．

TS 単位での識別は，一定の間隔で時間を区切り，その中に含まれるパケットを対象とする．フロー単位のようにアプリケーションごとにパケットを集約しないため，複数のアプリケーションの通信が TS 内に混在する場合がある．また TS 幅と TS の重複幅（スライド幅）の定義を調整する必要がある．

以下に，有効な TS 幅とスライド幅の検証実験を行った結果を述べる．

3.3. 識別アルゴリズム

TS 単位での識別は，スライド幅を小さくすると識別処理数が増大する．そのため，リアルタイム性を考慮した場合，高速に識別可能なアルゴリズムを用いることが好ましい．本稿では、検知率が高く，識別時間も短い特性を持つことが報告されている [2] SVM (Support Vector Machine)を用いることとした。

Study of the parameters extraction method for achieving the cyber attack detection

NantoSuzuki†, Airi Koike†, Takayuki Suzuki‡, and Noriharu Miyaho‡
†Graduate School of Information Environment, Tokyo Denki University
‡School of Information Environment, Tokyo Denki University

3.4. 特徴量の抽出

機械学習を行う際には，適切な特徴量の抽出が重要である．先行研究では，特徴量としてパケットサイズと到着間隔，レイヤ 4 以下のヘッダ情報，ペイロード情報などが，用いられている．本稿では，先行研究[3]より，異常通信検知に関して識別精度が高いと報告されている TS 内のパケットサイズの平均，TCP SYN パケットの割合，TCP ACK パケットの割合の 3 つの特徴量を用いることとした．

3.5. 実験方法

表 1 に実験諸元を示す．異常通信と通常通信を TS に分割し，TS から特徴量 3 種を抽出する．SVM により TS の正常・異常識別を行う（図 1）．

異常・正常通信の識別に有効な TS 幅とスライド幅の適切な値を抽出することを目的として実験評価を進めた．

表 1 実験諸元

データセット	正常： PC1 台の通信 4 時間分,ftp,メール (imap,pop,smtp),skype(チャット) 動画(YouTube,ニコニコ動画),p2p(torrent) 異常：CCCDataset 2011 年 1 月 28 日～31 日
識別アルゴリズム	SVM (LibSVM)
識別単位	タイムスロット（重複あり）
特徴量	タイムスロット内のパケットサイズ平均, ACK パケット割合,SYN パケット割合

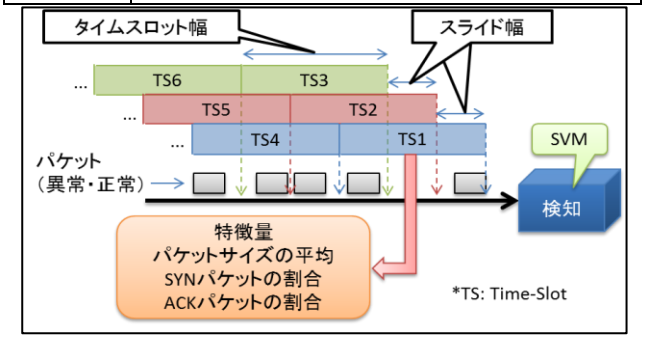

図 1 異常・正常通信データの検知実験

4. 実験結果と考察

図 2 に TS 幅を 1 秒から 10 秒に変化させ，スライド幅を 1 秒に固定した時の True Positive Rate (TP), True Negative Rate (TN)， Area Under The Curve (AUC)の関係を示す．TP は正常通信を正しく識別できた割合，TN は異常通信を正しく識別できた割合，AUC は識別器の精度評価を表す．TN は TS 幅 1 秒から 10 秒にかけて識別率が減少している．これは，異常通信は通信のパターンの変化間隔が短いため，TS 幅が小さい方が変化に対応し易いためと考えられる．TP は，TS 幅 1 秒から 3 秒まで上昇し，6 秒から 10 秒まで減少した．この結果から，正常通信の識別においては，一般的には 3～6 秒程度の TS 幅が妥当と推測できる．今回の実験では，長時間の動画通信や P2P 通信のデータを使用したため，TS 幅は異常通信に比べて大きい方が有効となったと判断できる．TS 幅が小さすぎると，他のアプリケーション通信の特徴と類似点が増えるため，誤識別が多くなり、結果として識別率が減少したと考えられる．一方，TS 幅が大きすぎると TS 内に複数の通信変化が混在し，特徴が平滑化されるため，識別率が減少したと考えられる．これらの結果から，異常通信も同様な識別傾向になる場合があることが推定できる．すなわち，異常通信の識別も TS 幅を 1 秒より小さくしていった場合に，ある時点で減少していくことが推測できる．図 3 に TS 幅を 10 秒に固定し，スライド幅を 1 秒から 10 秒に変化させた時の識別率との関係を示す．TN，TP ともにスライド幅 1 秒から 10 秒にかけて減少傾向を示す．この結果から，スライド幅は，小さい方が良い識別結果を得られると考えられる．

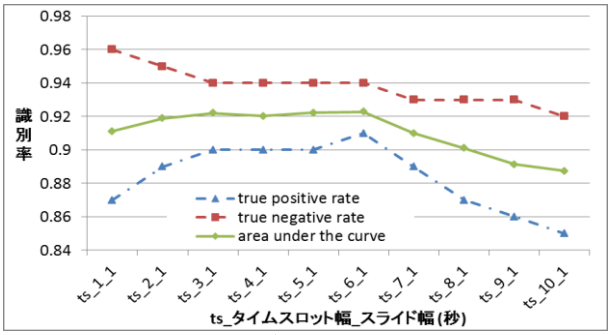

図 2 TS 幅の変化と識別率の関係

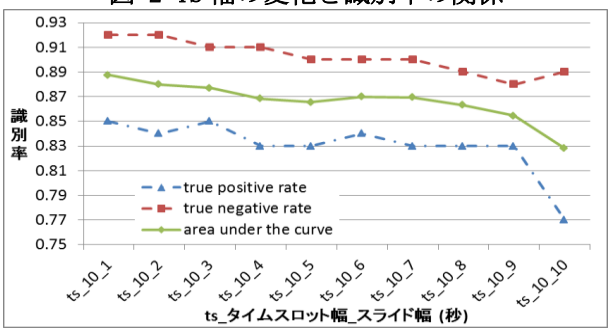

図 3 スライド幅の変化と識別率の関係

5. まとめと今後の課題

本稿では，正常通信と異常通信の識別実験を行い，TS 幅とスライド幅の変化による識別率の関係を調査した．実験結果から，正常通信に有効な TS 幅は通信内容によるが 3 秒から 6 秒となり，異常通信は 1 秒以下であることが推測できた．

今後は，通信内容ごとに TS 幅とスライド幅を調査する必要がある．

参考文献

[1] 神薗 雅紀 他: マルウェア対策のための研究用データセット ～MWS Datasets 2013～, CSS2013(MWS2013) (2013.10)

[2] S.Kondo and N.Sato. “Botnet traffic detection techniques by c&c session classification using svm,”IWSEC2007, Oct. 2007.

[3] 川元研治,市田達也,市野将嗣,畑田充弘,小松尚久, “マルウェア感染検知のための経年変化を考慮した特徴量評価に関する一考察”, MWS2011, (2011.10)