大分大学宛の HTTP 通信のクラスタリングによる特徴分析

大田尚吾 * 清水光司 ** 池部実 * 吉田和幸 ***
(大分大学 *工学部知能情報システム工学科 **工学研究科知能情報システム工学専攻 ***学術情報拠点情報基盤センター)

1 はじめに

DoS 攻撃やクロスサイトスクリプティング (XSS) などの Web サーバに対する攻撃が問題となっている．XSS により Web サーバに対して不正なプログラムを挿入された場合，そのサーバを閲覧したクライアントはマルウェアに感染する可能性がある．以上から，HTTP を利用した不正な通信への対策が必要である．

HTTP を利用した不正な通信の多くはプログラムによる自動化された通信である．そこで，HTTP 通信をユーザによる通信とプログラムによる通信に判別することで，HTTP 通信を利用した不正な通信の検知に役立つと考えた．本論文では大分大学宛の TCP/80 番ポートの通信 (HTTP 通信) について，クラスタリングにより分類する．

2 クラスタリングによる HTTP 通信の分類

2.1 HTTP 通信の分類

ユーザによる HTTP 通信は，ユーザが Web ブラウザを通じて Web ページへアクセスする通信である．また，プログラムによる HTTP 通信の 1 つに Web クローラがある．Web クローラは,Web 上で公開されているコンテンツを周期的に収集し，データベース化するプログラムである．主な Web クローラとして，Googlebot，bingbot，Baiduspider などがある．送信元のクローラの判定には，DNS 逆引き，HTTP リクエスト中の User-Agent などを用いる．本論文では，クローラとユーザによる通信を分類する．

2.2 データセット

分類対象の HTTP 通信として，インターネットから学内ネットワークの TCP/80 番宛のパケットを 2016 年 6 月 10 日 4 時 53 分から 2016 年 6 月 10 日 5 時 52 分の 1 時間，収集した．また，収集したデータに含まれる送信元を，DNS 逆引きにてクローラかどうか調べた．表 1 に分類対象のデータセットにおける TCP コネクション数，送信元ホスト数，クローラ数，宛先ホスト数を示す．

2.3 特徴ベクトルの定義

ユーザによる通信とプログラムによる通信を分類するために，3 つの特徴ベクトルを定義した．定義した特徴ベクトルを特徴ベクトル A，特徴ベクトル B，特徴ベクトル C と呼ぶ．また，マルウェアに感染したホストの場合，同一送信元からでもユーザによる通信とプログラムによる通信が発生する場合がある．そのため，特徴ベクトル A では TCP コネクション毎に HTTP 通信を分類する．特徴ベクトル B,C では送信元 IP アドレスをもとに分類する．

2.3.1 特徴ベクトル A

特徴ベクトル A では，TCP コネクション毎に以下の要素を用いて，TCP コネクション毎に分類する．(1)，(2) は，TCP ヘッダから得られる情報である．

(1) TCP コネクション毎の送信元の送信データ長

(2) TCP コネクション毎の送信元の受信データ長

(3) TCP コネクション毎の接続時間

表 1: データセット

コネクション数	送信元数	クローラ数	宛先数
5,417	501	162	97

2.3.2 特徴ベクトル B

特徴ベクトル B は，TCP コネクション毎に得られる以下の要素を用いて，送信元 IP アドレス毎に分類する．

(1) 送信元毎の送信元が送信した平均データ長

(2) 送信元毎の送信元が受信した平均データ長

(3) 送信元がアクセスした宛先ホスト数

(4) 送信元毎の平均 TCP コネクション接続時間

(5) 送信元毎の総 TCP コネクション数

2.3.3 特徴ベクトル C

特徴ベクトル C は片山の研究 [2] を参考に要素を決定した．[2] において，パケットに含まれるポート番号，プロトコル種類，TCP フラグを用いて分類していた．本研究では HTTP 通信を対象としているため，ポート番号など一部の要素を除外した．また，[2] において，HTTP 通信を分類するためには，別の要素を加える必要があるとのことなので，以下の (1),(2) の要素を加えた．

(1) 送信元が送受信したデータ長のうち，送信元が送信したデータ長の割合

(2) 送信元毎の平均 TCP コネクション接続時間

(3) パケット中の SYN フラグが立っている割合

(4) パケット中の PSH フラグが立っている割合

(5) パケット中の ACK フラグが立っている割合

(6) 送信元がアクセスした宛先ホスト数

2.4 クラスタリング

3 つの特徴ベクトルを用いて，クラスタリングによる分類をする．クラスタリングのアルゴリズムとして，k-means++[1] を使用した．k-means++は，分割最適化型クラスタリング手法の 1 つで，k-means 法の最初のクラスタ中心の選択に関して改良を加えたアルゴリズムである．今回は, 2 つのクラスタに分類する．各クラスタについて, クローラがどの程度含まれているかを検証する．クラスタリングには，Python の scikit-learn を用いて実行した．

3 実験結果と考察

表 2 に 3 つの特徴ベクトルを用いてクラスタリングで分類した結果を示す．

特徴ベクトル A では，コネクションのほとんどがクラスタ 1 に分類された．クラスタ 2 に分類されたコネクションはすべて同じ送信元であった．この送信元をホスト X と

表 2: クラスタリングによる分類結果

特徴ベクトル		コネクション数	送信元数	クローラ数
A	クラスタ 1	5,408	501	162
	クラスタ 2	9	1	0
B	クラスタ 1	2,641	491	158
	クラスタ 2	2,776	10	4
C	クラスタ 1	2,613	212	83
	クラスタ 2	2,804	289	79

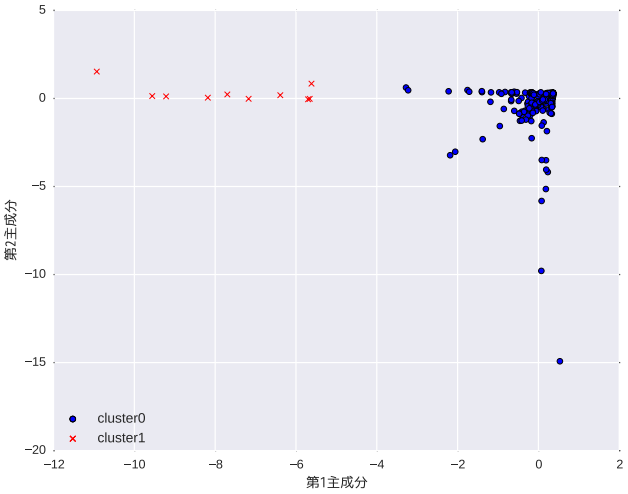

図 1: 特徴ベクトル B によるクラスタリングの分類結果

する．ホスト X のコネクションはクラスタ 1 にも含まれていた．表 3 にホスト X のコネクションの特徴ベクトル A の要素の値を示す．ホスト X の各クラスタでの要素の値に大きく差があった．同一送信元でも，その特徴から別のクラスタに分類できた．また，Web クローラのコネクションはすべてクラスタ 1 に分類された．しかし，特徴ベクトル A の分類ではホスト X の影響が大きいため，目的とする分類は達成できていない．

特徴ベクトル B によるクラスタリングの分類結果を図 1 に示す．特徴ベクトル B は，特徴ベクトル A の分類結果と同様に，多くのクローラがクラスタ 1 に分類された．しかし，表 2 に示すように，コネクション数がほぼ同じ割合で分類されていることから，クラスタ 1 に対してクラスタ 2 は送信元ごとのコネクション数が多い．

特徴ベクトル C によるクラスタリングの分類結果を図 2 に示す．特徴ベクトル C は，送信元とクローラ数がともに同じ割合で分類されていた．この結果から，特徴ベクトル C で定義した要素では目的とする分類結果を得ることはできなかった．

3.1 考察

結果として，3 つの特徴ベクトルでは目的とする分類結果を得ることはできなかった．特徴ベクトル A，B に関しては，両方の結果もデータ数の偏りがあったので，分割するクラスタ数を増やすことにより目的とする分類結果が得られると考察している．特徴ベクトル C では，さらに別の要素を加えて分類する必要がある．また，要素は，HTTP 通信の特性を十分に把握したうえで選択する必要がある．

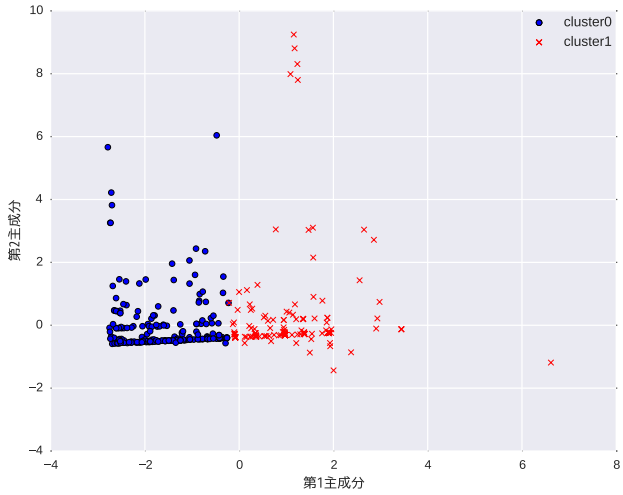

図 2: 特徴ベクトル C によるクラスタリングの分類結果

表 3: ホスト X における特徴ベクトル A での分類結果の各要素の値

(1)	(2)	(3)
クラスタ 1		
183	310	899.461
182	644	844.919
クラスタ 2		
190,886	148,556	4,551.651
190,868	148,561	4,779.903
190,819	147,461	4,864.485
190,911	148,556	4,995.278
190,864	147,461	5,638.459
190,773	148,561	5,580.340
190,835	147,461	6,187.023
191,011	148,561	6,011.560
190,807	147,461	5,238.354

4 まとめと今後の課題

本論文では，大分大学宛を HTTP 通信をユーザによる通信とプログラムによる通信に判別するために，3 つの特徴ベクトルを定義しクラスタリングによる分類をした．結果として，定義した 3 つの特徴ベクトルでは目的とする結果を得ることはできなかった．今後は，通信時間間隔を特徴ベクトルの要素に含めた関連研究 [3] などを参考に，特徴ベクトルに用いる要素を検討する．

参考文献

[1] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding”, Proceegings of the 18th Annual ACM SIAM Symposium on Discrete Algorithms(SODA’07), pp.1027–1035, 2007.

[2] 片山雄介, “クラスタリングによる通信の分類と時間変化の解析”, 早稲田大学 卒業論文, https://www.goto.info.waseda.ac.jp/forB4/pdf-th/2010/0203_k-yusuke.pdf, 2011.

[3] D. Ashley, “An Algorythm for HTTP bot dection”, University of Texas at Austin – Info. Security Office, Jan. 2011.