C07　Apriori と PrefixSpan による連携感染のハイブリッド検出方式

発表者 0BDRM004 大類 将之
指導教員 菊池 浩明 教授

Apriori–PrefixSpan Hybrid Approach for Automated Detection of Botnet Coordinated Attacks

Abstract: This paper aims to detect features of coordinated attacks by applying data mining techniques, Apriori and PrefixSpan, to the CCC DATAset 2008-2010 which consists of the captured packets data and the downloading logs. Data mining algorithms allow us to automate detecting characteristics from large amount of data, which the conventional heuristics could not apply. Apriori achives high recall but with false positive, while PrefixSpan has high precision but low recall. Hence, we propose hybriding these algorithms.

1　はじめに

近年のマルウェア (以下，MW) は，数多くの亜種が存在し，複数のダウンロード (以下，DL) サーバに分散して感染するなど，複雑化，高度化が進んでいる．特に分散されたサーバによる MW の連携感染は検出をより困難としている．そこで本研究では，研究用データセット CCC DATAset 2008~2010 [1] を用いて，データマイニング手法である Apriori と PrefixSpan を適用し，連携感染の特徴を報告する．また，2 つの手法を用いたハイブリッド検出方式を提案する．

2　要素技術

2.1　Apriori アルゴリズム [2]

Apriori は，Agrawal らが提案した代表的な頻出パターンマイニング手法である．支持度と確信度という閾値に最小値を与え，その値を元に数多く抽出される相関ルールの中から，効果的に価値の低いルールを枝刈りする．これにより，価値の低い相関ルールを除き，効率よく価値あるルールを発見できる．

2.2　PrefixSpan アルゴリズム [3]

PrefixSpan は，Pei らが提案した Prefix-projection という射影を用いた系列パターンマイニング手法である．系列データから，射影対象の系列より後ろに存在するアイテムのみを抽出し，深さ優先で射影を繰り返す事で，頻出する系列パターンを効果的に発見できる．

3　連携感染

3.1　定義

複数の DL サーバによる連携で，個別の MW を組み合わせて感染させる攻撃を連携感染と定義する．連携感染は起点となる MW を DL 後，IRC サーバに接続し，他の MW を HTTP GET により DL する．

3.2　実験データ

実験データとして，CCC DATAset 2008~2010 の攻撃通信データと攻撃元データを使用する．観測マシンは約 20 分間隔でクリーンな状態にリセットされるため，その間隔で各データを分割した．これをスロットと呼ぶ．スロットを 1 つのトランザクションとし，その間に DL された MW の種類をそのトランザクションに生じるアイテムとして，頻出アイテムを抽出する．

4　ハイブリッド検出方式

4.1　Apriori と PrefixSpan の比較

2009 年 2 月の実験結果の一部を表 1 に示す．これは，トレンドマイクロ社の報告にある `TSPY_KOLABC.CH`, `WORM_SWTYMLAI.CD`, `BKDR_POEBOT.GN` の連携感染を抜粋した結果である．Apriori では連携感染が検出されたスロットの頻度を “Slots”，PrefixSpan では検出された連携感染パターンの頻度を “Ptns” と定義する．それぞれ手動調査で定義した真のスロット，パターンの数を “True” とする．

表 1 の 2 月 4 日より，Apriori はスロット単位で確実に連携感染を含む 14 スロットを抽出している．しかし，2 月 28 日では 7 スロットの誤検出が確認できる．これは X と Y の違いから，実際には同じ相関ルールを別と認識して検出してしまうためである．一方，PrefixSpan では 4 種類の系列パターンをパターン単位で検出できている．しかし，3 種の MW による全ての考えられる組み合わせ 6 パターンのうち，`BKDR` を頭とするパターンは頻度が低いために除外され，検出できていない．以上の結果から，Apriori は連携感染を含むスロットの抽出，PrefixSpan は MW の正確な連携感染パターンの抽出に有効である．

Table 1: Apriori と PrefixSpan の比較

日付	Apriori			PrefixSpan		
	相関ルール (集合)	Slots	True [Slots]	系列パターン	Ptns	True [Ptns]
2009/02/03	WORM, BKDR ⇒ TSPY	4	4	TSPY ⇒ WORM ⇒ TKDR	3	9
2009/02/04	BKDR, TSPY ⇒ WORM	14	14	TSPY ⇒ BKDR ⇒ WORM	3	29
				TSPY ⇒ WORM ⇒ BKDR	7	
				WORM ⇒ BKDR ⇒ TSPY	4	
				WORM ⇒ TSPY ⇒ BKDR	12	
			⋮			
2009/02/28	BKDR, TSPY ⇒ WORM	7	7	TSPY ⇒ WORM ⇒ BKDR	5	14
	BKDR, WORM ⇒ TSPY	7		WORM ⇒ TSPY ⇒ BKDR	3	
合計		464	315		482	575

Table 2: 再現率と適合率

手法	再現率	適合率
Apriori	315/315 = 1.0	315/464 = 0.678
PrefixSpan	482/575 = 0.838	482/482 = 1.0
Hybrid	545/575 = 0.947	482/482 = 1.0

4.2 ハイブリッド検出方式

表 1 の 2 月 4 日より，Apriori では 1 種類のルール，PrefixSpan では 4 種類のパターンがあるが，実際には BKDR，TSPY，WORM の 3 種類の組み合わせだけでなく，その他の MW による連携感染も検出される．その数は Apriori が 4 種類，PrefixSpan が 32 種類であり，一見して連携感染の特定は難しい．そこで，Apriori と PrefixSpan を用いたハイブリッド検出方式を提案する．まず，Apriori を適用し，MW の組み合わせを特定する．次に，PrefixSpan を適用し，特定した MW に関するパターンのみを抽出する．最終的に 32 種類から 5 種類のパターンに限定し，その中から，頻度の高いパターンを選択して，それを連携感染として特定する．

5 実験結果

Apriori，PrefixSpan，提案方式それぞれの再現率，適合率を表 2 に示す．表 2 より，Apriori は再現率は高く，検出率は良いが，誤検出が発生する．PrefixSpan は適合率は高いが，未検出が発生するため，最適な最小支持度を設定する必要がある．提案方式は Apriori で先に連携感染を検出する事を前提とするため，PrefixSpan では枝刈りを考慮する必要がなくなり，支持度を一般的な値より低くできる．すなわち，PrefixSpan の支持度を下げて精度を求める事が可能となる．その結果として，未検出が減少し，再現率を向上できた．

6 おわりに

Apriori と PrefixSpan を用いた連携感染を自動検出するハイブリッド検出方式を提案した．提案方式は再現率が約 0.95，適合率が 1.0 と高い精度を示している．

参考文献

[1] 畑田，他，“マルウェア対策のための研究用データセット ∼MWS 2011 Datasets∼”，マルウェア対策研究人材育成ワークショップ 2011，pp. 1-5，2011.

[2] R. Agrawal, et al., “Mining Association Rules between Sets of Items in Large Databases”, Proc. of ACM SIGMOD-93, pp. 207-216,1993.

[3] J. Pei, et al., “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”, Proc. of the 17th Int’l Conf. on Data Engineering, pp. 215-224, 2001.

業績リスト

1. M. Ohrui, et al, “Mining Association Rules Consisting of Download Servers from Distributed Honeypot Observation”, The 13th Int’l Conf. on Network-Based Information Systems, pp. 541-545, 2010.
2. N. R. Rosyid, M. Ohrui, et al., “A Discovery of Sequential Attack Patterns of Malware in Botnets”, The 2010 IEEE Int’l Conf. on Systems, Man and Cybernetics, pp. 2564-2570, 2010.
3. M. Ohrui, et al., “Apriori-PrefixSpan Hybrid Approach for Automated Detection of Botnet Coordinated Attacks”, The 14th Int’l Conf. on Network-Based Information Systems, pp. 92-97, 2011.
4. N. R. Rosyid, M. Ohrui, et al., “Analysis on the Sequential Behavior of Malware Attacks”, IEICE Transactions on Information and Systems, Vol. E94-D, No. 11, pp. 2139-2149, 2011.

他 3 件.