УДК 004.056.5

*ГОЛОВІН Андрій Юрійович*

# ІДЕНТИФІКАЦІЯ ПРИХОВАНОЇ ПЕРЕДАЧІ ІНФОРМАЦІЇ У TCP/IP ТРАФІКУ НА ОСНОВІ ВИКОРИСТАННЯ МЕТОДУ НАЇВНОГО БАЄСІВСЬКОГО КЛАСИФІКАТОРА

**Постановка проблеми.** Прихована передача інформації є несанкціонованим процесом комунікації, що призводить до витоку інформації із контрольованої зони. Більшість систем виявлення та попередження вторгнень (IDS/IPS) можуть знаходити приховану передачу інформації у полях даних (payload) мережевих протоколів. Проте виявлення прихованої інформації у заголовках пакетів протоколу TCP/IP є непростим завданням. Складність цього процесу зумовлена випадковим характером значень деяких полів (наприклад, поля TCP ISN та IP ID). Для протидії прихованій передачі інформації із використанням вказаних полів необхідно відрізняти нормальні значення від тих, що використовуються для передачі інформації. Завдання встановлення приналежності елемента вибірки до певного класу вирішує метод наївного баєсівського класифікатора.

**Аналіз останніх досліджень і публікацій.** Методи виявлення прихованої передачі інформації та ідентифікації вторгнень у комп'ютерні мережі детально розглянуто у джерелах [1–5]. До основних методів ідентифікації прихованої передачі інформації відносять: аналіз сигнатури, статистичний та поведінковий аналіз, системи «процес-запит» (PQS), метод опорних векторів (Support Vector Machine, SVM).

SVM використовується в роботі [6] для виявлення прихованої передачі інформації в полях заголовка пакета протоколу TCP/IP: номер послідовності (ISN), прапорці (control flag) та контрольна сума (header checksum). В експерименті для навчання SVM використовувалися вибірки із 5000 нормальних та 5000 аномальних пакетів даних. Процес навчання є непростим, обчислювальна складність алгоритму SVM є високою.

У системах «процес-запит» (PQS) запити виражені як опис деякого процесу. Система запитів дозволяє PQS вирішувати нелегкі завдання з отримання даних. Наприклад, визначати процес за наявністю інформації про те, які вихідні дані генерує цей процес [3]. Ефективність методу у виявленні прихованої передачі інформації поки що є непідтвердженою, деякі моделі для опису процесів показують високий рівень хибних спрацювань, проте

автори публікації вважають, що метод є перспективним і підлягає подальшому аналізу.

**Метою** статті є поглиблене вивчення, розвиток і вдосконалення методів виявлення прихованої передачі інформації на основі використання методу наївного баєсівського класифікатора.

**Виклад основного матеріалу.** Значення полів TCP ISN та IP ID у заголовку (рис. 1) протоколу TCP/IP є випадковими величинами із рівномірним розподілом, проте функція генерації псевдовипадкових значень є відмінною для різних реалізацій стеку протоколів TCP/IP (рис. 2).

| Біти 0-3 | 4-7 | 8-15 | 16-18 | 19-23 | 24-31 |
|---|---|---|---|---|---|
| Версія | HLEN | Тип обслуговування | Загальна довжина | | |
| Ідентифікація | | | Прапорці | Зміщення фрагментації | |
| Час життя | | Протокол | Контрольна сума заголовку | | |
| IP-адреса відправника | | | | | |
| IP-адреса отримувача | | | | | |
| Опції | | | | | Додаток |
| Дані (65535 мінус заголовок) | | | | | |
| ... | | | | | |

| Біт | 0 — 3 | 4 — 9 | 10 — 15 | 16 — 31 |
|---|---|---|---|---|
| 0 | Порт джерела | | | Порт призначення |
| 32 | Номер послідовності | | | |
| 64 | Номер підтвердження | | | |
| 96 | Зсув даних | Зарезервовано | Прапорці | Вікно |
| 128 | Контрольна сума | | | Вказівник важливості |
| 160 | Опції (необов'язково) | | | |
| 160/192+ | Дані | | | |

Рис. 1. Поля IP ID та TCP ISN в заголовку пакета (виділені прямокутниками)

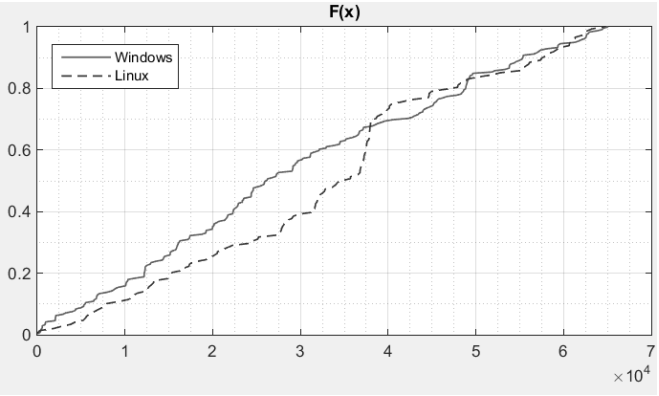



Рис. 2. Графік функції розподілу значень IP ID для ОС Windows та Linux

Наївний баєсівський класифікатор (Naive Bayes Classifier, NBC) – класифікатор, що ґрунтується на теоремі Баєса:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)},$$

де

$P(c|d)$ – імовірність, що значення $d$ належить до класу $c$;
$P(d|c)$ – імовірність наявності значення $d$ серед усіх значень класу $c$;
$P(c)$ – безумовна імовірність наявності даних класу $c$ у вибірці;
$P(d)$ – безумовна імовірність наявності значення $d$ у вибірці.

Мета класифікації полягає у визначенні класу приналежності для значення із вибірки. Для цього використовується не безпосереднє значення імовірності, а найбільш ймовірний клас. Баєсівський класифікатор використовує оцінку апостеріорного максимуму *(Maximum a posteriori estimation)* для визначення найбільш ймовірного класу (класу із максимальною ймовірністю):

$$c_{map} = \arg\max_{c \in C} \frac{P(d|c)P(c)}{P(d)}. \quad (1)$$

Оскільки безумовна імовірність наявності значення $d$ у вибірці $P(d)$ є константою, формула (1) буде мати вигляд:

$$c_{map} = \arg\max_{c \in C} [P(d|c)P(c)]. \quad (2)$$

Із врахуванням припущення про умовну незалежність (імовірність значень із вибірки не залежать одне від одного), а також проблеми арифметичного переповнення, визначення найбільш ймовірного класу буде мати вигляд:

$$c_{map} = \arg\max_{c \in C} \left[ \log P(c) + \sum_{i=1}^{n} \log P(w_i|c) \right].$$

Оцінка імовірності $P(c)$ та $P(w_i|c)$ виконується на навчальній вибірці. Імовірність класу оцінюється як:

$$P(c) = \frac{D_c}{D},$$

де

$D_c$ – кількість значень, що належать класу $c$;

$D$ – загальна кількість значень у навчальній вибірці.

Оцінка імовірності значення в класі (на прикладі мультиноміальної моделі):

$$P(w_i|c) = \frac{w_{ic}}{\sum_{\hat{i} \in V} w_{\hat{i}c}},$$

де

$W_{ic}$ – кількість разів, коли $i$-те значення зустрічається в класі $c$;

$V$ – словник корпусу вибірки (список всіх унікальних значень).

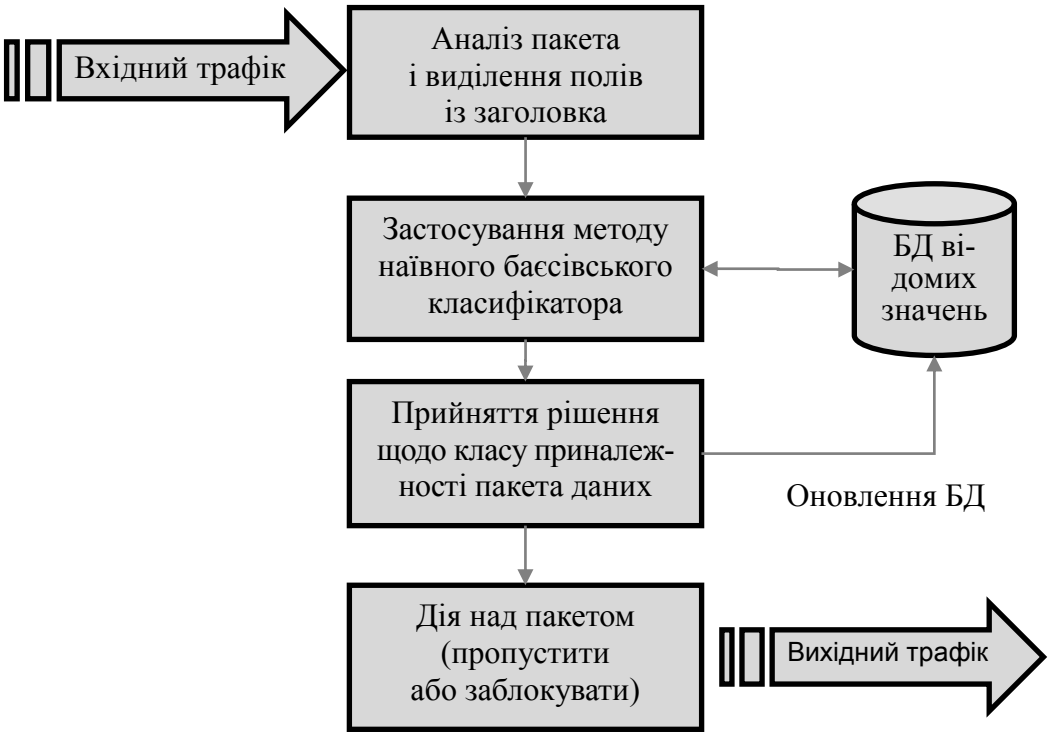
Із застосуванням адитивного згладжування (для вирішення проблеми невідомих змінних) кінцевий вираз, згідно якого буде виконуватись класифікація, має вигляд:

$$c_{map} = \arg\max_{c \in C} \left[ \log \frac{D_c}{D} + \sum_{i=1}^{n} log \frac{w_{ic}+1}{|V|+\sum_{\hat{i} \in V} w_{\hat{i}c}} \right]. \quad (3)$$

Використовуючи формулу (3) для визначення приналежності значення із вибірки необхідно розрахувати $c_{map}$ для обох класів: присутня прихована передача, відсутня прихована передача (легітимний трафік). Порівнявши розраховані значення $c_{map}$, можна зробити висновок про приналежність значення до конкретного класу.

Перевагою цього підходу є те, що вимоги до розміру вибірки скорочуються від експоненційних до лінійних. Недоліком є те, що модель точна лише у випадку, коли виконується припущення про умовну незалежність. В іншому випадку обчислені ймовірності вже не є точними (і навіть більше того, їх сума може не дорівнювати одиниці, через що потрібно нормувати результат). Однак на практиці невеликі відхилення від незалежності призводять лише до незначного зниження точності, і навіть у разі істотної залежності між змінними результат роботи класифікатора продовжує корелювати з істинною приналежністю образу до класів. При цьому переваги використання класифікатора (висока швидкість роботи, простота і масштабність, помірні вимоги до пам’яті) часто переважують недоліки.

***Модель системи ідентифікації прихованої передачі інформації***




У теорії використання наївного баєсівського класифікатора забезпечує вищу швидкість навчання (тренування), ніж інші існуючі методи машинного навчання. Запропонований підхід використання методу NBC полягає у створенні аналітичної моделі для ідентифікації прихованої передачі інформації у заголовку пакета TCP/IP, яка матиме більш високу швидкість навчання, ніж будь-який інший існуючий підхід. За допомогою методу NB можна реалізувати програмний класифікатор значень полів IP ID та TCP ISN для розпізнавання значень, що використовуються для прихованої передачі інформації.

**Висновки.** У статті розглянуто теоретичні основи роботи наївного баєсівського класифікатора, переваги та недоліки його застосування. Також запропоновано модель системи ідентифікації прихованої передачі інформації із використанням методу наївного баєсівського класифікатора. Для визначення ефективності застосування зазначеного методу необхідно реалізувати програмний компонент, що використовує класифікації TCP/IP пакетів за методом NBC. Окремо треба визначити вимоги (пам’ять, обчислювальні ресурси, час) для процесу тренування системи.

## Список використаних джерел

1. Головін А. Ю. Методи виявлення прихованих каналів передачі інформації у комп'ютерних мережах / А. Ю. Головін // Збірник наукових праць ІПМЕ. – 2014. – Вип. 71. – С. 9.

2. Cybenko G., Berk V., Crespi V., Robert S. Gray, Jiang G. An Overview of Process Query Systems. – 2004.

3. Cybenko G., Berk V., Crespi V., Robert S. Gray, Jiang G. Covert Channel Detection Using Process Query Systems. – 2005.

4. Tumoian E., Anikeev M. Detecting NUSHU Covert Channels Using Neural Networks . – 2005.

5. Sohn T., Jung Seo J., Moon J. A Study on the Covert Channel Detection of TCP/IP Header Using Support Vector Machine. – 2003.

6. T. Sohn, J. S., J. Moon, «A study on covert channel detection of TCP/IP header using support vector machine» in Proc. 5th Int. Conf. Information and Communication Security (ICICS2003), pp. 313–324, Oct. 2003.

7. Apurva N. Mahajan, I. R. Shaikh, «Review on covert channel detection methods of TCP/IP header», International Journal of Computer Science International Journal of Computer Sciences and Engineering, vol. 2. – 2014.

8. Vibhor Kumar Vishnoi, Sunil Kumar, «Detection of TCP/IP Covert Channel based on Naive Bayesian Classifier», International Journal Of Engineering And Computer Science Volume. – 3 Issue. – 9. – 2014.

*Рецензенти:*
кандидат технічних наук Ю. Іващенко,
кандидат технічних наук Н. Григоренко

***Аннотация:*** В статье рассмотрен подход идентификации скрытой передачи информации в TCP/IP трафике (поле TCP ISN и поле IP ID в заголовках TCP/IP пакета) на основе использования метода наивного баесовского классификатора.

***Ключевые слова:*** скрытая передача информации, скрытый канал TCP/IP, наивный баесовский классификатор.

***Abstract:*** Article focuses on detection methods of TCP/IP covert channel (using TCP ISN and IP ID fields in TCP/IP packet header) based on Naive-Bayesian Classifier.

***Key words:*** covert channel, TCP/IP, Naive-Bayesian classifier.