УДК 004.056.5

ПОРІВНЯННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПОШУКУ ШКІДЛИВОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Мєтєлєва Л.В.[1]

[1] *Національний технічний університет України "Київський політехнічний інститут", Фізико-технічний інститут*

Анотація

У роботі розглянуто проблему класифікації шкідливого програмного забезбечення у високонавантажених системах за допомогою алгоритмів машинного навчення.

Ключові слова: алгоритми класифікації, машинне навчення, шкідливе програмне забезпечення, високонавантежені системи

Вступ

У високонавантажених систем є свої особливості, що вимагають відмінного від інших комп'ютерних систем підходу, у тому числі й під час класифікації шкідливого програмного забезпечення та виявлення аномалій у бінарних файлах. Для таких систем є недопустимим виділення значних ресурсів для запуску підозрілих файлів у "пісочниці", потокової перевірки виконуваних файлів антивірусом чи для використання інших методів поведінкового аналізу.

1. Методика досліджень

Поширені методи поведінкового аналізу, такі як імітація запуску в реальній системі та пошук аномальної поведінки, є доволі ефективними, але вимагають значних обчислювальних ресурсів. Для розв'язання поставленої проблеми можна використовувати статистичні методи. Але для цього необхідне використання значної за обсягом бази сигнатур, яку потрібно постійно оновлювати. А також, статистичні методи не дозволяють виявити файли інфіковані новими видами шкідливого програмного забезпечення. Для того щоб позбутися цих недоліків, застосуємо метод машинного навчання. Це узагальнена назва штучної генерації знань з досвіду. Штучна система навчається на прикладах і після закінчення фази навчання може самостійно приймати рішення. Тобто система не просто порівнює підозрілі дані з відомими зразками, як у статистичних алгоритмів, а розпізнає певні закономірності в навчальних даних. Найбільш ефективними сучасними алгоритмами машинного навчання є нейронні мережі, J48, PART, SVM та інші. Зручним інструментом для проведення класифікації є набір засобів візуалізації та алгоритмів для аналізу даних і вирішення задач прогнозування - Weka. Weka дозволяє виконувати такі завдання аналізу даних, як підготовку даних (preprocessing), відбір ознак (feature selection), кластеризацію, класифікацію, регресійний аналіз та візуалізацію результатів.

2. Результати дослідження

В результаті досліджень було створено на мові програмування Python оптимальний алгоритм на основі аналізу публічних баз зразків шкідливого коду. Для достатньої кількості зразків для здійсненна аналізу бінарних послідовностей було обрано наступні ресурси: VirusShare, vxheaven.org та malwr.com. Також окремо формується вибірка файлів без шкідливого навантаження, для попередження помилок другого роду, тобто коли "нешкідливий" файл буде позначено як "шкідливий". Він може аналізувати бінарні виконувані файли Win32 (EXE або DLL), поділені на три групи: "чисті", "інфіковані" та "невідомі". Для тестування роботи програми використовувалася вибірка з 42000 зразків шкідливого програмного забезпечення та 2000 000 "чистих" файлів. У PE заголовках виконуваних файлів відокремлено такі атрибути:

- NumberOfSections
- SizeOfCode
- SizeOfInitializedData
- SizeOfUninitializedData
- AddressOfEntryPoint
- BaseOfCode
- BaseOfData
- ImageBase
- SizeOfImage
- SizeOfHeaders
- SizeOfStackReserve
- SizeOfStackCommit
- SizeOfHeapReserve
- SizeOfHeapCommit
- NumberOfRvaAndSizes
- ImageVersion

- IatRVA
- DebugSize
- ExportSize
- ResourceSize

У ході роботи дані отриманні з заголовку невідомого файлу аналізуються і відповідно до результату аналізу приймається рішення щодо категорії даного файлу. Було обрано такі алгоритми машинного навчання: SVM, J48 Graft та PART.

Табл. 1. Попередні результати на тестовій вибірці

Алгоритм	Вірно класифіковані		Невірно класифіковані	
J48	43899	99.809%	84	0.191%
SVM	43297	98.4403 %	686	1.5597 %
PART	43905	99.8227 %	78	0.1773 %

Крім цього, для підвищення ефективності уже відомих алгоритмів можна використати метод бустингу. Це процедура послідовної побудови композиції алгоритмів машинного навчання, коли кожен наступний алгоритм намагається компенсувати недоліки композиції всіх попередніх алгоритмів. Під час доповіді будуть наведені результати експериментального дослідження ефективності вище загаданих алгоритмів класифікації шкідливого програмного забезпечення.

Висновок

Запропоновано статистичний метод класифікації на основі алгоритмів машинного та методів бустингу для застосування в високонавантажених системах мережевої фільтрації.

Література

1. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition). Ian H. Witten, Eibe Frank and Mark A. Hall— 2012. 664 p.
2. Sumeet Dua, Xian Du.Data Mining and Machine Learning in Cybersecurity — Auerbach Pub, 2010. — 240 p.
3. Marcus A. Maloof.Machine Learning and Data Mining for Computer Security — Springer Science & Business Media, 2006. — 210 p.