

ШВИДКОДІЯ МОДЕЛІ NAÏVE BAYES З ВІДБОРОМ ОЗНАК ДЛЯ НАБОРУ USB-IDS-1

Штанько В.І. аспірант, науковий керівник Нікітенко Є.В.

Системи виявлення атак (IDS) відіграють ключову роль у сфері кібербезпеки. Модель Naïve Bayes при цьому відзначається простою реалізацією та швидкою обробкою даних. Водночас її продуктивність істотно залежить від правильного вибору ознак. Тому дослідження впливу кількості ознак на точність класифікації та ресурсні витрати є надзвичайно актуальним, зокрема для систем, що функціонують у режимі реального часу.

Модель для навчання була побудована наступним чином: спочатку дані навчального набору USB-IDS-1 [1] були очищені від ознак з порожніми значеннями, а також від ознак із малою варіативністю, оскільки вони ускладнюють класифікацію даних з допомогою моделі Naïve Bayes. В якості алгоритму для вибору ознак було використано PCA (Principal component analysis) [2, 3]. Кількість ознак зменшувалась із 74 (максимальна кількість для обраного набору даних) до 10. Для кожної ітерації циклу було виконано по 20 повторень, для кожного з яких обраховувався час навчання і тестування моделі, а також такі метрики як точність (accuracy), влучність (precision), повнота (recall) [4] та оцінка F1. В якості навчальних даних були використані дані з набору USB-IDS-1, а саме дані про атаки типу TCPFlood, Slowloris, Hulk та Slowhttptets з файлів NoDefence та Evasive, які були відмічені як атаки. Також було використано дані про нормативний трафік [1].

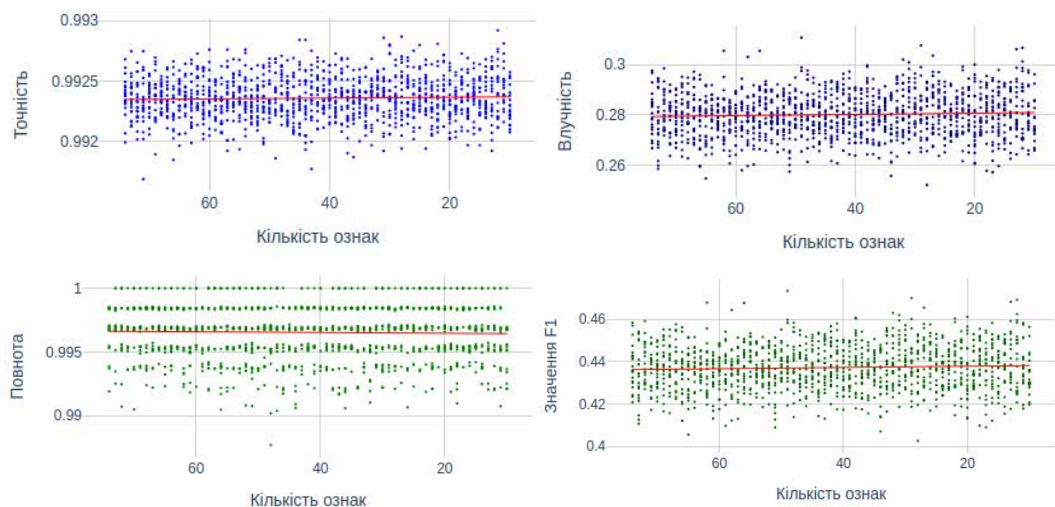


Рисунок 1 – Залежність оцінок ефективності моделі від кількості ознак

Результати дослідження показали наступні закономірності. Значення точності залишалися стабільним на рівні ~99.2% незалежно від кількості використаних ознак. Це свідчить про загальну надійність моделі в класифікації. Влучність також є фактично незмінним значенням для будь-якої кількості ознак в зазначеному діапазоні, і незначно коливається в межах від 26% до 30%. Показник повноти теж залишався дуже високим (> 99.5%) у всіх конфігураціях, що вказує на здатність моделі ефективно виявляти всі випадки атак, не залишаючи значної кількості хибно-негативних результатів. Як гармонійне середнє між влучністю та повнотою, F1 теж незначно коливається в межах 42 - 46%, тобто не має чітких ознак зростання чи спадання залежно від кількості ознак (рис. 1).

У ході дослідження оцінювалась залежність часу виконання моделі Naïve Bayes від кількості ознак, які поступово вилучались із повного набору (74 ознаки) до мінімального (10 ознак). Результати показали чітку майже лінійну залежність: при 74 ознаках час

обробки одного прикладу становив у середньому близько 850 мс, тоді як при 10 — лише близько 150 мс. Зменшення кожних приблизно 10 ознак призводило до скорочення часу на 100–120 мс (рис. 2). Незначні відхилення від лінійного тренду можна пояснити тим, які саме ознаки видалялися — деякі з них, ймовірно, мали більшу обчислювальну складність, наприклад, через числовий тип чи широкий діапазон значень.

Таке спостереження має важливе прикладне значення: оптимізація кількості ознак дозволяє суттєво покращити швидкість моделі без втрати якості класифікації. У системах реального часу, де обробка має відбуватись з мінімальною затримкою, наприклад у міжмережних екранах або моніторингу мережевого трафіку, скорочення обчислювального навантаження є критичним. На основі результатів можна стверджувати, що вилучення понад 50% ознак не погіршує метрик точності, recall чи F1, але забезпечує п'ятикратне зменшення часу роботи. Таким чином, попередній аналіз і відбір ознак є не лише доцільним, а й необхідним кроком при практичному застосуванні Naïve Bayes у задачах кібербезпеки.

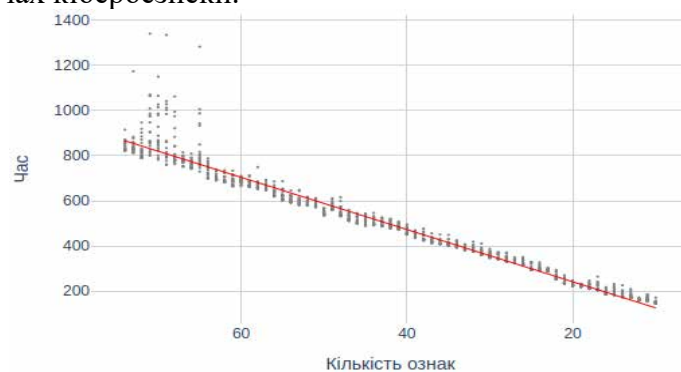


Рисунок 2 – Залежність тривалості навчального циклу від кількості ознак

Для підвищення ефективності моделі Naïve Bayes у задачах виявлення атак доцільно використовувати попередній відбір ознак, орієнтуючись на найбільш інформативні та уникаючи надлишкових або корельованих, які можуть негативно впливати на точність класифікації та порушувати припущення моделі. Скорочення кількості ознак дозволяє суттєво зменшити час обробки без втрати якості. Рекомендовано адаптувати набір ознак до конкретного середовища, тестуючи його ефективність на репрезентативних або реальних даних.

Таким чином було встановлено, що для набору USB-IDS-1 зменшення кількості ознак у моделі Naïve Bayes не призводить до суттєвих змін якості класифікації, але значно зменшує час обробки даних. Це підкреслює важливість використання методів відбору ознак для оптимізації продуктивності моделей, особливо в задачах реального часу, таких як виявлення кіберзагроз.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Università Degli Studi del Sannio. USB-IDS Datasets. (2022). [Online]. Available: <https://idsdata.ding.unisannio.it/datasets.html>
2. Zhang J. (2019). Machine Learning With Feature Selection Using Principal Component Analysis for Malware Detection: A Case Study. Sophos Technical Papers, <https://doi.org/10.48550/arXiv.1902.03639>
3. Labib, Khaled & Vemuri, Rao. (2006). An application of principal component analysis to the detection and visualization of computer network attacks. *Annales des Télécommunications*. 61. 218-234. <http://dx.doi.org/10.1007/BF03219975>
4. Gushin, I.V., Sych, D.O. (2018). Analysis of the Impact of Text Preprocessing on the Results of Text Classification. *Molodyi vchenyi*, 10, 264-266. <https://molodyivchenyi.ua/index.php/journal/article/view/3755>

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І
ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**



**ТЕОРЕТИЧНІ ТА ПРИКЛАДНІ
АСПЕКТИ РОЗРОБКИ
КОМП'ЮТЕРНИХ СИСТЕМ**

**ЗБІРНИК НАУКОВИХ ПРАЦЬ ЗА МАТЕРІАЛАМИ
VII ВСЕУКРАЇНСЬКОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ
СТУДЕНТІВ І АСПІРАНТІВ
*24 квітня 2025 року***

Київ 2025

УДК 004

Відповідальна за випуск: М.І. Лендел

Збірник наукових праць за матеріалами VII Всеукраїнської науково-практичної конференції студентів і аспірантів «ТЕОРЕТИЧНІ ТА ПРИКЛАДНІ АСПЕКТИ РОЗРОБКИ КОМП'ЮТЕРНИХ СИСТЕМ '2025», 24 квітня 2025 року, НУБіП України, Київ. – 452 с. (електронне видання)

Відповідальність за зміст публікацій несуть автори.

Передрук матеріалів, а також використання їх будь-якій формі допускається лише з дозволу авторів