АНАЛІЗ ВПЛИВУ МЕТОДІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ НА ЕФЕКТИВНІСТЬ МОДЕЛЕЙ ВИЯВЛЕННЯ КІБЕРЗАГРОЗ

Фісенко Д.М., Мазепа А.Д., Пантєлєєв В.О.

Кафедра інфокомунікаційної інженерії ім. В.В. Поповського,
Харківський національний університет радіоелектроніки,
Україна.

E-mail: dmytro.fisenko@nure.ua
artem.mazepa@nure.ua
vadym.pantielieiev@nure.ua

Abstract

This article investigates how various data preprocessing techniques impact ML model performance designed for cyber threat detection. The study evaluated how data cleaning, normalization, feature transformation, and class balancing affect the probability of reliable detection of malicious activity. The analysis revealed that the selection and combination of preprocessing methods can improve the quality of datasets, which can improve model performance. These results helped us justify the choice of specific preprocessing strategies to ensure a given level of detection accuracy and robustness. Based on the results it is possible create a recomendations for selecting the best data preprocessing approaches. These recommendations may depend on the type of cyber threats, how clean the raw data is, and the reliability requirements of threat detection systems.

В сучасному світі системи виявлення кіберзагроз в хмарних середовищах це складні комплекси, що поєднують мережеві екрани, сервери обробки даних, та системи виявлення вторгнень. Однією з ключових характеристик таких систем є ефективність моделей машинного навчання, які використовуються для виявлення кіберзагроз. На результат роботи моделей машинного навчання значною мірою впливає якість наборів даних (датасетів), які використовуються у процесі їх тренування та експлуатації. Якість даних характеризує здатність сформованого набору даних підтримувати необхідний рівень ефективності моделі в умовах впливу різноманітних факторів. Це поняття охоплює властивості повноти даних, їхньої чистоти, інформативності ознак, збалансованості класів, а також стійкості до шумів та нестандартних записів. Повнота забезпечує наявність достатньої кількості прикладів як нормальної активності, так і різних типів атак. Чистота даних – відсутність некоректних та зашумлених записів. Інформативність ознак – відповідність полів журналів ключовим поведінковим характеристикам. Збалансованість – рівномірність кількості прикладів різних класів. Стійкість до шумів – відсутність значного впливу випадково спотворених чи продубльованих записів на модель машинного навчання.

В процесі аналізу сучасних датасетів, було виявлено що основними причинами зниження ефективності моделей є неповні або некоректні значення, велика кількість пропусків, продубльовані записи, зашумлений мережевий трафік, нерівномірність розподілу між «атаками» і «нормальною» активністю, а також неконсистентні значення ознак. Ці фактори суттєво ускладнюють побудову точних і стабільних моделей, особливо в середовищах з високою варіативністю кіберзагроз. Забезпечення високої якості даних потребує злагодженої роботи більшості методів попередньої обробки таких як очищення, нормалізації, кодування категоріальних ознак, вибору релевантних параметрів і балансування класів. Кожен із цих етапів відіграє критичну роль. Очищення даних дозволяє видалити пропуски, дублікатні або пошкодженні записи. Нормалізація забезпечує коректне функціонування алгоритмів, чутливих до масштабу ознак. Кодування категоріальних атрибутів (протокол, порт, тип події) уніфікує представлення даних для моделей. Балансування класів суттєво підвищує здатність виявляти рідкісні та складні типи атак. Вибір ознак сприяє зменшенню надлишковості та покращує узагальнювальні властивості моделей [1].

Очищення даних – це початковий і один із найважливіших етапів попередньої обробки даних. Він забезпечує якість набору даних для подальшого використання. Мета очищення полягає у видаленні або виправленні некоректних, неповних, продубльованих або зашумлених записів, які можуть негативно впливати на якість моделей машинного навчання, їхню точність, стабільність та узагальнювальну здатність. Пропуски у даних виникають унаслідок технічних обмежень мережевого обладнання, втрати пакетів, неповної фіксації подій або перевантаження систем логування. Наявність пропусків призводить до некоректного навчання моделей машинного навчання або до відмови алгоритмів від обробки таких записів. Для обробки пропущених значень застосовуються такі методи:

– видалення записів із великою кількістю пропусків, якщо їх частка у датасеті незначна;
– заповнення числових пропусків середнім, медіаною або спеціальними маркерами;
– заповнення категоріальних пропусків іншими значенням або найпоширенішою категорією.

Зазвичай дублікати виникають під час агрегування записів, збору даних різними джерелами або реєстрації однієї й тієї самої події IDS системою декілька разів. Їх наявність порушує рівномірність даних і модель може «переучуватися» на повторюваних даних, що призводить до зростання похибки узагальнення. Для виявлення дублікатів виконують порівняння ключових полів або використовують хеш-ідентифікацію подій. Повні дублікати видаляються, а часткові можуть агрегуватися. Некоректні або неможливі значення є типовими для мережевих логів. Наприклад записи з від'ємними значеннями, некоректними портами, недійсними IP адресами чи неконсистентними часовими мітками. Такі записи можуть призвести до помилок під час навчання або створити хибні закономірності. Для їх обробки застосовують:

– видалення записів із явними помилками;
– заміну неможливих значень спеціальними маркерами;
– корекцію форматів (наприклад, переведення часу в UTC).

Якісно проведене очищення даних суттєво покращує навчання моделей машинного навчання. Завдяки цьому зменшується кількість хибних спрацювань, підвищуються показники Recall та F1-score. Також слід зазначити що, моделі набувають більшої стійкості до варіацій у даних, що прискорюється процес навчання за рахунок зменшення обсягу зайвих даних [2], [3], [4].

Наступним етапом попередньої обробки даних є нормалізація числових ознак. Він забезпечує узгодженість масштабів параметрів та сприяє стабільному і точному навчанню моделей машинного навчання. У задачах виявлення кіберзагроз ця процедура набуває особливої важливості, оскільки мережеві характеристики можуть мати різні діапазони та розподіли. Нормалізація дозволяє уникнути домінування ознак із великими числовими значеннями над тими, що мають менший порядок, забезпечуючи рівний внесок кожної ознаки в процес навчання моделі.

Також значущим етапом є кодування категоріальних ознак. З його допомогою можливо виконати адаптацію символьних або номінальних параметрів до формату, придатного для математичної обробки алгоритмами машинного навчання. Для більшості моделей машинного навчання числове представлення ознак є обов'язковим, оскільки саме числові значення можуть бути піддані операціям порівняння, обчислення відстаней, градієнтів або побудови дерев рішень. У задачах виявлення кіберзагроз категоріальні ознаки займають значну частку інформації, наприклад тип протоколу, напрямок трафіку, тип події в журналі тощо. Тому коректне кодування таких даних суттєво впливає на точність і якість моделей.

В процесі навчання моделі машинного навчання важливу роль відіграє вибір релевантних ознак. Від правильності вибору параметрів залежить точність, стабільність та швидкість роботи моделі. У контексті виявлення кіберзагроз вибір ознак відіграє особливо важливу роль через специфіку мережевих даних. Вони можуть мати великі обсяги, високу розмірність, наявність дублюючих або неінформативних параметрів, а також нерівномірний розподіл класів. Неправильно підібрані ознаки можуть призвести до перенавчання, збільшення обчислювальних витрат і зниження здатності моделі виявляти складні або рідкісні види атак [5], [6].

Балансування класів є також важливим етапом попередньої обробки даних, оскільки у реальних мережевих журналах класи розподіляються нерівномірно. Більшість записів представляють нормальну активність, тоді як інциденти інформаційної безпеки або аномалії становлять лише невелику частку загального обсягу даних. Це призводить до дисбалансу класів, який суттєво ускладнює навчання моделей.

Моделі машинного навчання, навчені на дисбалансних даних, схильні ігнорувати рідкісні події та прогнозують переважно нормальні (безпечні) події. Це призводить до високої загальної точності,

але дуже низької здатності до виявлення атак. Для систем виявлення вторгнень це є критичним недоліком, оскільки навіть одна пропущена атака може мати значні наслідки для безпеки. Тому балансування класів є необхідною умовою для отримання точних, надійних і стійких моделей для виявлення кіберзагроз [7].

У ході проведеного дослідження було проаналізовано вплив основних методів попередньої обробки даних таких як очищення, нормалізації числових ознак, кодування категоріальних параметрів, вибору релевантних ознак та балансування класів. Також розглянуто їх вплив на ефективність роботи моделей машинного навчання у задачах виявлення кіберзагроз. Результати дослідження підтверджують, що якість вхідних даних є критичним фактором, який безпосередньо визначає точність, надійність і стійкість алгоритмів у системах виявлення вторгнень. Очищення усуває пропуски, дублікатні та некоректні записи, зменшуючи шум і ризик перенавчання. Нормалізація забезпечує коректну роботу моделей, чутливих до масштабу, приводячи числові ознаки до узгоджених діапазонів. Кодування категоріальних ознак дозволяє трансформувати символьні параметри у числову форму без втрати семантики. Вибір релевантних ознак зменшує розмірність і підвищує точність за рахунок виключення малозначущих параметрів. Балансування класів вирішує проблему переважання нормального трафіку над зловмисним і суттєво підвищує показники виявлення рідкісних загроз. У сукупності ці методи здатні покращити ефективність моделей на 20 – 30%, що робить їх критичним етапом у побудові сучасних систем кіберзахисту.

*Л*ітература

1. Dhongade, Gauri & Chandrakar, Dr & Khande, Rajeshree. (2024). Enhancing Cyber Security : A Study of Data Preprocessing Techniques for Cyber Security Datasets. International Journal of Scientific Research in Science and Technology. 11. 71-75. 10.32628/IJSRST2411427.

2. Manocchio, Liam & Layeghy, Siamak & Gallagher, Marcus & Portmann, Marius. (2025). An empirical evaluation of preprocessing methods for machine learning based network intrusion detection systems. Engineering Applications of Artificial Intelligence. 158. 111289. 10.1016/j.engappai.2025.111289.

3. Songma, Surasit & Sathuphan, Theera & Pamutha, Thanakorn. (2023). Optimizing Intrusion Detection Systems: Exploring the Impact of Feature Selection, Normalization and Three-Phase Precision on the Cse-Cic-Ids-2018 Dataset. 10.20944/preprints202311.0302.v1.

4. Ketepalli, Gayatri & Bulla, Premamayudu. (2023). Data Preparation and Pre-processing of Intrusion Detection Datasets using Machine Learning. 257-262. 10.1109/ICICT57646.2023.10134025.

5. Kalimuthan, C. & Renjit, J.. (2020). Review on intrusion detection using feature selection with machine learning techniques. Materials Today: Proceedings. 33. 10.1016/j.matpr.2020.06.218.

6. Uwazie, Emmanuel & Obiniyi, Afolayan & Olalere, Morufu. (2024). Handling Class Imbalance in Intrusion Detection Dataset using Undersampling and Oversampling Techniques (2024). 17. 24-33.

7. Shanmugam, Vaishnavi & Razavi-Far, Roozbeh & Hallaji, Ehsan. (2024). Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches. Electronics. 14. 69. 10.3390/electronics14010069.