Оптимізація порогу прийняття рішення в системах IDS/IPS на основі моделей машинного навчання

УДК 004.8:004.056

Каріна Крушельницька[1], Дмитро Тимощук[2], Наталія Загородна[3]

Тернопільський національний технічний університет імені Івана Пулюя,
[1]karina.kryshel@gmail.com, [2]dmytro.tymoshchuk@gmail.com,
[3]zagorodna_n@tntu.edu.ua

Цифровізація, розширення мережевої інфраструктури та зростання обсягів мережевого трафіку супроводжуються підвищенням кількості й складності кіберзагроз, що зумовлює активне використання методів машинного навчання в системах IDS/IPS. На відміну від сигнатурних підходів, які переважно ґрунтуються на відомих шаблонах атак і тому мають обмежену ефективність щодо нових загроз, алгоритми машинного навчання здатні аналізувати поведінкові ознаки трафіку та виявляти аномалії, характерні для нових або модифікованих атак, зокрема атак «нульового дня».

ML-класифікатори зазвичай формують не лише бінарну мітку класу, а й числову оцінку ймовірності або впевненості моделі, яку порівнюють із порогом класифікації τ. Без попереднього калібрування такі оцінки можуть бути зміщеними, особливо для дерев рішень, ансамблевих і бустингових моделей, що ускладнює прямий вибір порогу за шкалою ймовірностей. Хоча значення τ = 0,5 часто використовується як стандартне, у задачах IDS/IPS воно не завжди є оптимальним через незбалансованість класів і різну вартість помилок класифікації [1]. Для отримання більш достовірних імовірнісних оцінок перед вибором порогу доцільно застосовувати методи калібрування, а оптимальне значення τ визначати за метриками, релевантними до задачі виявлення вторгнень. Вибір порогу класифікації слід розглядати не лише як технічний параметр ML-моделі, а як важливе проєктне рішення, що визначає баланс між рівнем виявлення атак, кількістю хибних спрацювань і загальною ефективністю системи захисту.

Якість класифікації в задачах IDS/IPS доцільно оцінювати на основі матриці помилок, елементами якої є TP, TN, FP та FN. За умови, що позитивним класом вважається атака, TP відповідає правильно виявленій атаці, TN — правильно класифікованому нормальному трафіку, FP — хибній тривозі, коли легітимний трафік помилково визначено як атаку, а FN — пропущеній атаці. На основі цих величин обчислюють основні метрики оцінювання ефективності ML-класифікатора, наведені в таблиці 1.

Таблиця 1

Основні метрики оцінювання ML-класифікатора в задачах IDS/IPS

Метрика	*Формула*	*Що вимірює*	*Особливості застосування в IDS/IPS*
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Частку всіх правильних класифікацій	Може бути неінформативною при дисбалансі класів, оскільки високе значення може досягатися за рахунок домінування нормального трафіку

Precision	$\frac{TP}{TP+FP}$	Частку справжніх атак серед усіх зразків, класифікованих як атаки	Важлива для зменшення хибних тривог, особливо в IPS, де FP може спричинити блокування легітимного трафіку
Recall/TPR /Sensitivity	$\frac{TP}{TP+FN}$	Частку реально наявних атак, які були виявлені моделлю	Важлива для IDS/IPS, оскільки низьке значення Recall означає велику кількість пропущених атак
F1-score	$\frac{2 * Precision * Recall}{Precision + Recall}$	Гармонійне середнє між Precision і Recall	Доцільна як компромісна метрика, коли потрібно збалансувати хибні тривоги та пропущені атаки
ROC-AUC	Площа під ROC-кривою	Загальну здатність моделі розділяти класи при різних порогах	Може переоцінювати якість моделі при сильному дисбалансі класів
PR-AUC	Площа під Precision–Recall-кривою	Якість виявлення позитивного класу за різних порогів	Особливо корисна при дисбалансі класів, коли атаки становлять меншість
FPR	$\frac{FP}{FP+TN}$	Частку нормального трафіку, помилково класифікованого як атака	Критична для IPS, оскільки високий FPR може призводити до блокування легітимного трафіку та перевантаження системи реагування
TNR/Specificity	$\frac{TN}{TN+FP}$	Частку нормального трафіку, правильно класифікованого як нормальний	Важлива для оцінювання здатності системи не створювати хибних тривог; особливо актуальна для IPS, де FP може призводити до блокування легітимного трафіку
G-Mean	$\sqrt{TPR \cdot TNR}$	Збалансованість виявлення атак і правильного розпізнавання нормального трафіку	Корисна при дисбалансі класів, оскільки враховує якість класифікації як позитивного, так і негативного класу

Таким чином, у задачах IDS/IPS недостатньо оцінювати модель лише за показником Accuracy, оскільки ця метрика може бути неінформативною за умов дисбалансу класів. Більш обґрунтованим є комплексне використання Precision, TPR, TNR, F1-score, PR-AUC, FPR та G-Mean, що дає змогу оцінити баланс між здатністю системи виявляти атаки, рівнем хибних спрацювань і якістю розпізнавання нормального трафіку [2]. Оптимальне значення порогу класифікації доцільно визначати не довільно, а на основі цільового критерію, що відповідає функціональному призначенню IDS/IPS. Залежно від пріоритетів системи таким критерієм може бути максимізація F1-score, G-Mean або статистики Youden's J (J = TPR − FPR), досягнення заданого балансу між Precision і Recall, мінімізація очікуваної вартості помилок класифікації або максимізація Recall за умови допустимого рівня FPR.

Отже, універсального значення порогу τ не існує. Його вибір має залежати від операційної ролі системи, допустимого рівня ризику та вартості помилок класифікації. Для IDS пріоритетним може бути підвищення Recall з метою мінімізації пропущених атак, тоді як для IPS особливо важливо контролювати FPR, оскільки хибні спрацювання можуть призводити до блокування легітимного трафіку та порушення доступності сервісів. Тому поріг класифікації доцільно розглядати як керований проєктний параметр, що визначає практичну ефективність ML-моделі в конкретному середовищі розгортання.

1. Tymoshchuk, D., Zagorodna, N., Klots, Y., Yatskiv, V., Petliak, N. AutoML and explainable AI-based approach to enhance the efficiency and interpretability of IDS. CEUR Workshop Proceedings, 2025, 4163, pp. 231-246
2. Tymoshchuk, D., Sverstiuk, A., Klots, Y., Petliak, N., Titova, V. An explainable artificial intelligence approach for detecting network attacks. CEUR Workshop Proceedings, 2025, 4141, pp. 38-51

Застосування нечітких продукційних правил для контекстно-довірчого оцінювання кіберризиків у середовищі Інтернету речей

УДК 621.395.7 (043.2)

Підлісний Юрій[1], Шелест Михайло[2]

Національний університет «Чернігівська політехніка», [1]ypodlesny@ukr.net

Стрімкий розвиток Інтернету речей (IoT) призвів до масового впровадження інтелектуальних пристроїв у різних сферах, що супроводжується зростанням кіберзагроз через обмежені ресурси та спрощені механізми захисту [1].

Традиційні методики оцінювання ризиків орієнтовані на класичні інформаційні системи та недостатньо ефективні в IoT через динамічність середовища, гетерогенність пристроїв та неповноту даних [2]. У таких умовах доцільним є застосування методів нечіткої логіки, які дозволяють працювати з експертними оцінками та невизначеними параметрами [3].

У роботі запропоновано підхід до оцінювання кіберризиків у середовищі IoT на основі нечітких продукційних правил типу IF–THEN (табл.1). Наведена база правил є фрагментом знань, який може бути розширений або адаптований залежно від специфіки IoT-середовища та моделі загроз.

Таблиця 1

Фрагмент бази нечітких продукційних правил

№	Правило	Результат
1	IF V = High AND T = High AND D = Low THEN R = Critical	Критичний
2	IF V = Medium AND T = High THEN R = High	Високий
3	IF V = Low AND T = Medium AND D = High THEN R = Medium	Середній
4	IF S = High AND V = Low THEN R = Low	Низький
5	IF A = High AND T = Medium THEN R = High	Високий
6	IF C = High AND V = Medium THEN R = High	Високий
7	IF S = Low AND T = High THEN R = Critical	Критичний

Запропонована модель враховує технічні характеристики вузлів, стан середовища та достовірність даних, що забезпечує її застосування як у статичних, так і в динамічних IoT-системах.

Вхідними параметрами моделі є: рівень вразливості (V), інтенсивність загроз (T), рівень захищеності (S), критичність активу (C), мережева аномальність (A) та довіра до джерела даних (D). Вихідною змінною є інтегральний показник ризику R, що характеризує ступінь небезпеки для конкретного вузла, підсистеми або сегмента мережі.