

ВДОСКОНАЛЕННЯ МЕТОДУ ISOLATION FOREST ДЛЯ ВИЯВЛЕННЯ ІНСАЙДЕРСЬКИХ ЗАГРОЗ У CRM- СИСТЕМАХ

Вінницький національний технічний університет

Анотація

У роботі розглянуто проблему автоматизованого виявлення інсайдерських загроз у CRM-системах, що оперують гетерогенними даними. Запропоновано модифікацію алгоритму Isolation Forest шляхом введення механізму змішаних розбиттів (Mixed Splits), який дозволяє обробляти категоріальні ознаки без застосування One-Hot кодування. Експериментально підтверджено підвищення точності детекції та зменшення використання оперативної пам'яті.

Ключові слова: інсайдерські загрози, Isolation Forest, CRM, змішані дані, аномалії.

Abstract

The paper considers the problem of automated insider threat detection in CRM systems operating with heterogeneous data. A modification of the Isolation Forest algorithm is proposed by introducing a Mixed Splits mechanism, which allows processing categorical features without One-Hot encoding. Increased detection accuracy and reduced RAM usage were experimentally confirmed.

Key words: insider threats, Isolation Forest, CRM, mixed data, anomalies.

Вступ

Сучасні системи управління взаємовідносинами з клієнтами (CRM) акумулюють критичні бізнес-дані, що робить їх пріоритетною ціллю не лише для зовнішніх атак, але й для внутрішніх порушників. Специфіка інсайдерських загроз полягає у використанні легітимного доступу для нелегітимних цілей (масовий експорт бази, махінації з цінами). Головною проблемою існуючих методів машинного навчання при аналізі логів CRM є гетерогенна природа даних, де числові параметри (сума транзакцій) співіснують з категоріальними (роль користувача, тип операції). Класичні методи адаптації даних, такі як One-Hot Encoding, призводять до розрідженості простору ознак та зниження ефективності алгоритмів, що обумовлює актуальність розробки спеціалізованих підходів.

Результати дослідження

Базовим методом для виявлення аномалій обрано алгоритм Isolation Forest (iForest), який базується на принципі ізоляції рідкісних подій. Однак стандартна реалізація iForest працює виключно з числовими даними, використовуючи порогові розбиття вигляду $x < p$. Для обробки категоріальних даних зазвичай застосовують кодування, що штучно збільшує розмірність простору та нівелює семантичні зв'язки між категоріями.

У роботі запропоновано модифікацію алгоритму, названу Mixed Isolation Forest. Ключовою відмінністю є введення адаптивної логіки розбиття вузлів дерева. Якщо обрана для спліту ознака є числовою, застосовується стандартний пороговий метод. Якщо ж ознака категоріальна, використовується метод випадкових підмножин (Subset Split): множина унікальних категорій ділиться на дві непусті підмножини S та S' , а критерієм переходу по дереву стає умова належності

значення до підмножини $X \in S$.

Такий підхід дозволяє працювати з "сирими" даними, усуваючи проблему "прокляття розмірності". Для верифікації методу було проведено серію експериментів на реальному датасеті Olist E-Commerce Dataset (понад 100 тис. записів), до якого було ін'єктовано синтетичні аномалії, що моделюють фінансові махінації та логістичне шахрайство.

Порівняльний аналіз зі стандартною реалізацією (iForest + One-Hot Encoding) показав, що запропонований метод демонструє кращу здатність до кластеризації контекстних аномалій. Зокрема, метрика ROC-AUC зросла з 0.945 до 0.998. Важливим технічним результатом стало зменшення споживання оперативної пам'яті у 2.3 рази, що пояснюється відсутністю необхідності зберігати розріджену матрицю ознак.

Висновки

Запропонований метод Mixed Isolation Forest довів свою ефективність для задач моніторингу безпеки в CRM-системах. Відмова від попереднього кодування категоріальних ознак дозволила не лише підвищити точність виявлення інсайдерських загроз, але й суттєво оптимізувати обчислювальні витрати. Це робить метод перспективним для впровадження в системи реального часу, де критичними є швидкість реакції та ресурсомісткість.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Liu F. T., Ting K. M., Zhou Z. H. Isolation forest // 2008 Eighth IEEE International Conference on Data Mining. – IEEE, 2008. – P. 413-422.
2. Hariri S., Kind M. C., Brunner R. J. Extended isolation forest // IEEE Transactions on Knowledge and Data Engineering. – 2019. – Vol. 33. – No. 4. – P. 1479-1489.
3. Вахній Д.Д. Вдосконалення алгоритму Isolation Forest шляхом введення змішаних сплітів у задачі виявлення інсайдерських загроз у CRM-системах: магістерська кваліфікаційна робота спец. 125 "Кібербезпека та захист інформації" / Д.Д. Вахній – Вінниця: ВНТУ, 2025. – 95 с.

Вахній Денис Дмитрович – студент групи 2KITS-24м, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, Вінниця.

Науковий керівник: **Салієва Олена Володимирівна** – д.ф., доцент кафедри менеджменту та безпеки інформаційних систем, Вінницький національний технічний університет, Вінниця.

Vakhniy Denys D. – student of group 2KITS-24m, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia.

Supervisor: **Saliieva Olena V.** – Doctor of Philosophy, Associate Professor of the Department of Management and Security of Information Systems, Vinnytsia National Technical University, Vinnytsia.