УДК 004.93’14

магістр О.О. Маркова, к.т.н., доцент Т.А. Желдак
Державний ВНЗ «Національний гірничий університет»

УДОСКОНАЛЕННЯ АЛГОРИТМУ ІДЕНТИФІКАЦІЇ НЕБАЖАНОГО ВТОРГНЕННЯ В ЛОКАЛЬНУ МЕРЕЖУ В УМОВАХ НЕДОСТАТНЬОЇ ІНФОРМАЦІЇ ІЗ ЗАСТОСУВАННЯМ ТЕХНОЛОГІЙ OLAP

В роботі обґрунтовано застосування методів класифікації видів з’єднань, що встановлюються локальною комп’ютерною мережею із глобальною мережею на “нормальні” та “небезпечні”. Також систематизовано множину ознак потенційно небезпечних комп’ютерних з’єднань для ідентифікації типу несанкціонованого вторгнення. Як результат - покращено ефективність алгоритму класифікації типів небезпечних комп’ютерних з’єднань для запобігання несанкціонованого вторгнення з використанням технологій OLAP.

Програмне забезпечення для ідентифікації вторгнень захищає комп’ютерну мережу від несанкціонованого доступу, незалежно, виконується таке вторгнення ззовні чи за допомогою так званих інсайдерів (співробітників чи клієнтів, що мають доступ до мережі зсередини). Однією з основних задач, що вирішуються при створенні програмного забезпечення для захисту комп’ютерних мереж є створення класифікатора з високим ступенем надійності, який би в реальному масштабі часу розділяв з’єднання на «нормальні» та «небажані», а також ідентифікував би тип вторгнення чи нападу.

В роботі пропонується для розв’язання задачі навчання датчика несанкціонованого вторгнення застосувати технології OLAP (On-Line Analytical Processing, з англійської – аналітична обробка в реальному часі). Вперше подібний підхід було запропоновано Американською асоціацією сприяння обороні DARPA у 2000 році [1], коли до масиву з даних про з'єднання за допомогою методології KDD (Knowledge Discovery in Databases) було застосовано алгоритми швидкої класифікації.

В якості вихідних даних були використані дані про з’єднання локальних комп’ютерних мереж ВПС США за 9 тижнів 1999 року. Масив складається з близько 5 мільйонів записів навчальної вибірки (перші сім тижнів роботи мережі) та близько 2 мільйонів записів тестової вибірки (останні два тижні

роботи). Інформація про кожне з'єднання – рядок з описом послідовності пакетів TCP, класифікованих за 42 ознаками. Запис про кожне з'єднання дорівнює приблизно 100 байтам, тож загальний обсяг вибірки сягає 670 Мб.

Як показало дослідження, всі несанкціоновані вторгнення поділяються на чотири категорії:

– атаки DOS (відмова в обслуговуванні), наприклад, повінь SYN і т.і.;

– R2L (remote-to-local) – несанкціонований доступ з віддаленого комп'ютера, наприклад, злам пароля;

– U2R (user-to-root) – несанкціонований доступ до локального користувача з виключними правами адміністратора («кореневого» облікового запису);

– зондування – спостереження за роботою мережі й пошук її слабких місць, наприклад, сканування відкритих і незахищених логічних портів.

Таким чином, розглянута задача перетворилася на три етапну: захисне програмне забезпечення, отримавши відомості про нове з'єднання, має, насамперед, визначити, чи є воно «нормальним», чи «небажаним», потім - ідентифікувати тип загрози і нарешті – конкретизувати вид атаки чи вторгнення. останніх було розглянуто 24 у навчальній виборці та 38 – у тестовій.

Застосування OLAP для розв'язання поставленої задачі, найперше, дозволяє застосовувати пошук у великих базах даних у реальному масштабі часу. Для обробки навчальної вибірки було створено куб даних, де метриками стали середні значення, кількості певних ознак, а також суми по тим чи іншим вимірам.

Було з'ясовано, що різні вимірювання мають різну інформативну цінність для класифікації видів з'єднання. В ході аналізу було вирішено відмовитись від 6 з 42 ознак через їх не інформативність (відсутність зв'язку між ознакою та видом з'єднання), та ще чотирьох – через тавтологію (повна кореляція між двома різними ознаками).

Серед 32 ознак, що залишилися, більшість (21) – безперервні значення, решта (11) – дискретні та логічні.

Основним механізмом класифікатора було обрано так звану наївну мережу Байєса (Naive Bayesian), у кожному з вузлів якої у якості логічного елемента застосовується множинна логістична регресія. Остання дозволяє створювати нечітко-логічну структуру висновку без втрати безперервного характеру значень ознак, за якими виконується класифікація. Згадана нечітка логічна мережа була реалізована у середовищі MATLAB.

У якості альтернативного рішення було розглянуто дерево рішень, побудова якого виконувалася за допомогою алгоритму C4.5 у середовищі Deductor Acedemic. В ході реалізації було застосовано відсікання усіх правил, що зустрічаються менше 4 разів та мінімальний рівень достовірності у 20%.

В ході експериментальної перевірки роботи запропонованих алгоритмів було підтверджено життєздатність обох підходів. Слід відзначити значний час, який потребує алгоритм навчання класифікатора на основі Байєсової мережі з логістичною регресією, оскільки розрахунок коефіцієнтів регресії виконується методом найменших квадратів. Однак навчання класифікатора відбувається одноразово у режимі офф-лайн, тоді як розпізнавання кожного окремого мережевого з'єднання виконується «на льоту». Тому значні часові витрати для навчання класифікатора не є його суттєвим недоліком.

Порівняння отриманих результатів засвідчує більшу точність класифікації наївними Байєсовими мережами у порівнянні з деревами рішень як при поділі на «нормальні» та «небажані» з'єднання (98,6% вірних відповідей проти 94%), так і при ідентифікації конкретного типу мережевої атаки (96,1% проти 89,2%). Водночас, обидва алгоритми показали стійку помилку ідентифікації типів вторгнень, що зустрічалися в навчальній виборці менше 1% разів (4 та 6 типів відповідно).

Перелік літератури:

1. Stolfo S. J. "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project" / Salvatore J. Stolfo, Wenke Lee, Wei Fan, Andreas

Prodromidis, and Philip K. Chan // DARPA Information Survivability Conference. – 2000.