УДК 004.891.3

Т.М. Долінський, А.М. Стефанів
Тернопільський національний технічний університет імені Івана Пулюя, Україна

МЕТОДИ ВИЗНАЧЕННЯ ФІШИНГУ ІЗ ЗАСТОСУВАННЯМ ТЕХНОЛОГІЇ APACHE SPARK MLLIB

T.M. Dolinskii, A.M. Stefaniv

PHISHING DETECTION METHODS USING APACHE SPARK MLLIB

Сьогодні, комп'ютерні технології стрімко розвиваються та використовуються у всіх сферах повсякденного життя. Це часто використовується зловмисниками, котрі фальсифікують оригінальні джерела такі як: сторінки входу в системи банкінгу, засоби переведення коштів, інтернет магазини з метою обману користувача та викрадення його персональних даних, ця проблема отримала назву фішинг.

З метою захисту користувача, у сфері безпеки інформаційних технологій, існують способи запобігання фішингу. У цій області досить ефективно використовуються засоби машинного навчання, котрі базуються на алгоритмах аналізу даних, в рамках даної публікації розглянуто алгоритми логістичної регресії та дерева рішень.

Логістична регресія є широко використовуваним методом завдяки своїм легко інтерпретованим і практичним результатам. Ця модель є ефективною для прогнозування двійкових даних, оскільки вона спирається на статистичні дані і застосовує узагальнену лінійну модель. Незважаючи на простоту цього методу, вона має три недоліки: по-перше, він вимагає більше статистичних допущень до застосування. По-друге, метод більш залежить від змінних, які мають лінійний зв'язок, ніж ті, що мають складне відношення. Нарешті, точність прогнозування є чутливою до повноти даних.

Дерево рішень - це графічна модель класифікації, яка складається з вузлів і країв. Базовий вузол називається Root, з якого починається дерево рішень. Кожен вузол всередині мережі містить правило "Якщо-то", клас і функцію, а також веде до наступного, використовуючи стрілки, які називаються краями. Дерева рішень закінчується вузлом листів, який називається термінатором. Дерево може включати в себе один або декілька етапів класифікаторів та внутрішні вузли, обмежені кореневими та кінцевими вузлами Ці методи мають багато реалізацій на різних мовах програмування. В рамках дослідження, для демонстрації використання та порівняння реалізації алгоритмів, використано засоби Apache Spark MLlib та Python Sklearn.

У вирішенні проблеми виявлення фішингу був використаний типовий процес тестування машинного навчання. Спочатку була надана функція всіх даних з набору даних. Наступним кроком було навчання моделі з набором тестів, який містить 70% вихідних даних, для максимального прогнозування. На останньому кроці було відправлено останні 30% даних для перевірки.

Значення точності, з параметрами за замовчуванням, реалізації Apache Spark MLlib, отримане під час експерименту було краще, ніж результати реалізації Python Sklearn: 92% проти 90%. Це означає, що технологія Apache Spark має кращу реалізацію алгоритмів і може допомогти забезпечити кращу масштабованість та швидкодійність в нових системах для аналізу фішингових даних. Реалізація розподілу даних в Apache Spark MLlib дає змогу розпаралелювати навчання з мільйонами або навіть мільярдам екземплярів.