ZAŁĄCZNIK NR 2
do Wniosku o wszczęcie postępowania w sprawie nadania stopnia doktora

Dissertation abstract

This dissertation presents a method for detecting adversarial attacks on machine learning models that classify tabular data. The method operates in a black-box regime, requiring only access to the model's inputs and outputs, without knowledge of its architecture or parameters.

The research was motivated by an observed security gap in systems using machine learning models for decision-making processes. This gap stems from not accounting for ML models' vulnerabilities to adversarial attacks, particularly those using minimal but targeted modifications of input data. Additional motivation came from the lack of developed methods for detecting adversarial attacks in black-box scenarios and the underrepresentation of research on attacks targeting models processing tabular data.

The proposed method is based on two key components: a surrogate model approximating the behavior of the diagnosed model, and diagnostic attributes extracted from this surrogate model reflecting the model's behavior in local neighborhoods of analyzed examples. Based on these attributes, a trained classifier detects attack cases within a given data window.

The experimental work included comprehensive testing on 22 datasets, using three types of machine learning models (logistic regression, SVM, XGBoost) and three different attack methods (ZOO, HopSkipJump, PermuteAttack). The research confirmed the statistical correlation between diagnostic attribute values and attack occurrence, and demonstrated the effectiveness of classifiers in detecting attacks (balanced accuracy 0.87-0.97). The method's ability to detect new types of attacks was verified by successfully identifying three previously unseen attacks (BIM, PGD, FGM) targeting a new type of machine learning model (neural networks), confirming the method's generalization capabilities.

The limitations of the method include its unsuitability for distinguishing between attack types and its current design focusing only on tabular data classification models. The method is also limited to detecting attacks at the inference stage and assumes uncompromised model operation during the preparation of the surrogate model.

The research was conducted as part of an implementation doctorate in collaboration with QED Software, which specializes in developing machine learning model explainability techniques. The results will be used in the company's platform for monitoring and continuous auditing of ML models.

This work contributes to the field of machine learning security by providing a practical approach to detecting adversarial attacks while operating under realistic constraints of limited access to the monitored model. The method's ability to detect previously unseen types of attacks makes it particularly valuable for real-world applications.

Streszczenie rozprawy doktorskiej w języku polskim

W niniejszej rozprawie doktorskiej przedstawiono metodę wykrywania ataków adwersaryjnych na modele uczenia maszynowego klasyfikujące dane tabelaryczne. Metoda działa w reżimie czarnej skrzynki, wymagając jedynie dostępu do danych wejściowych i wyjściowych modelu, bez znajomości jego architektury czy parametrów.

Badania zostały zainspirowane zaobserwowaną luką w zabezpieczeniach systemów wykorzystujących modele uczenia maszynowego w procesach decyzyjnych. Luka ta wynika z nieuwzględniania podatności modeli ML na ataki adwersaryjne, w szczególności te wykorzystujące minimalne, ale celowe modyfikacje danych wejściowych. Dodatkową motywację stanowił brak rozwiniętych metod wykrywania ataków adwersaryjnych w scenariuszach czarnej skrzynki oraz niedoreprezentowanie badań nad atakami na modele przetwarzające dane tabelaryczne.

Zaproponowana metoda opiera się na dwóch kluczowych komponentach: modelu zastępczym aproksymującym działanie diagnozowanego modelu oraz atrybutach diagnostycznych wydobywanych z tego modelu zastępczego, odzwierciedlających zachowanie modelu w lokalnych otoczeniach analizowanych przykładów. Na podstawie tych atrybutów, wytrenowany klasyfikator wykrywa przypadki ataku dla zadanego okna danych.

Prace eksperymentalne objęły kompleksowe testy na 22 zbiorach danych, z wykorzystaniem trzech typów modeli uczenia maszynowego (regresja logistyczna, SVM, XGBoost) oraz trzech różnych metod ataku (ZOO, HopSkipJump, PermuteAttack). Badania potwierdziły statystyczną korelację między wartościami atrybutów diagnostycznych a występowaniem ataku oraz wykazały skuteczność klasyfikatorów w wykrywaniu ataków (balanced accuracy 0,87-0,97). Zdolność metody do wykrywania nowych typów ataków została zweryfikowana poprzez skuteczną identyfikację trzech wcześniej nieznanych ataków (BIM, PGD, FGM) skierowanych na nowy typ modelu uczenia maszynowego (sieci neuronowe), potwierdzając zdolności generalizacyjne metody.

Ograniczenia metody obejmują jej nieprzydatność do rozróżniania typów ataków oraz obecne ukierunkowanie wyłącznie na modele klasyfikacji danych tabelarycznych. Metoda jest również ograniczona do wykrywania ataków na etapie inferencji i zakłada niezaburzone działanie modelu podczas przygotowywania modelu zastępczego.

Badania zostały przeprowadzone w ramach doktoratu wdrożeniowego realizowanego we współpracy z firmą QED Software, specjalizującą się w rozwoju technik wyjaśnialności modeli uczenia maszynowego. Wyniki zostaną wykorzystane w platformie firmy służącej do monitorowania i ciągłego audytu modeli ML.

Praca wnosi wkład w dziedzinę bezpieczeństwa uczenia maszynowego, dostarczając praktyczne podejście do wykrywania ataków adwersaryjnych przy jednoczesnym działaniu w realistycznych warunkach ograniczonego dostępu do monitorowanego modelu. Zdolność metody do wykrywania wcześniej nieznanych typów ataków czyni ją szczególnie wartościową dla zastosowań rzeczywistych.