

# 1. Temat projektu

Nienadzorowana detekcja anomalii za lasu izolacyjnego. Funkcje do tworzenia modelu i predykcji. Porównanie z nadzorowaną detekcją anomalii za pomocą dostępnych w R algorytmów klasyfikacji.

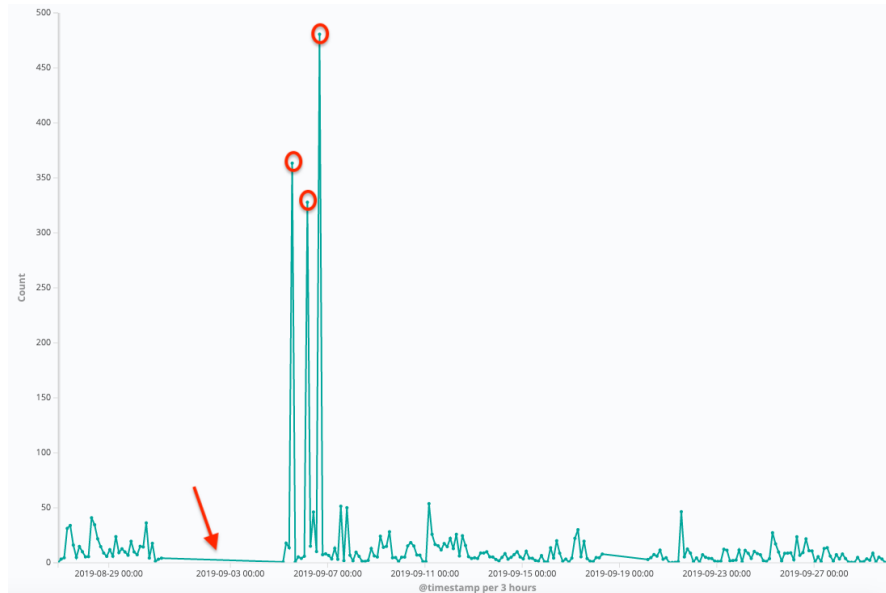
## 2. Skład zespołu

- Zhan Banzekulivakha (317683)
- Adam Krawczyk (317678)

## 3. Wstęp teoretyczny

**Las izolacyjny** - to nienadzorowany algorytm uczenia się do wykrywania anomalii, który działa na zasadzie izolowania anomalii, po raz pierwszy zaproponowany w 2008 r. [1], a opublikowany w 2012 [2].

Pod słowem anomalia rozumiemy dane, które przedstawia obserwację lub zdarzenie, które mają inną charakterystykę niż pozostałe. Na przykład: wykres na rysunku 1 przedstawia ruch wejściowy do serwera WWW. Wystarczy spojrzeć na obrazek, że niektóre punkty (zaznaczone czerwonym kółkiem) są niezwykle wysokie, do tego stopnia, że można podejrzewać, iż serwer WWW mógł być w tym czasie atakowany. Z drugiej strony płaski segment wskazany czerwoną strzałką również wydaje się nietypowy i może być oznaką, że serwer był wyłączony w tym okresie [3].



Rysunek 1. Przykładowy ruch internetowy z potencjalnymi anomaliami [3].

Większość istniejących podejść do wykrywania anomalii są metody oparte na klasyfikacji bądź to: K najbliższych sąsiadów (angl. *k-nearest neighbors algorithm*), klasyfikacja jednoklasowa (angl. *One-Class SVM*), metody oparte na klastrach (angl. *Clustering-Based methods*), Sieć Neuronowa Replikatora (angl. *Replicator Neural Network*) itp.. Jednak te algorytmy zwykle mają tzn. "side-effect" lub pierwotnie były wykorzystywane do

innych celów niż wykrywanie anomalii (np. do grupowania albo klasyfikacji). Prowadzi to do dwóch głównych wad [2]:

- 1) Istniejące podejście nie są zoptymalizowany pod kątem wykrywania anomalii , co oznacza że będziemy mieć styczeń z tzw. fałszywymi anomaliami (normalnie obserwacji/zdarzenia oznaczone jako anomalia) lub zbyt małą liczbą wykrytych anomalii
- 2) Wiele istniejących metod przeznaczone są do pracy z małą liczbą danych.

Isolation Forest - to algorytm który początkowo był zaprojektowany do wykrywania anomalii i stosuje inne podejście: zamiast próbować zbudować model normalnych instancji, wyraźnie izoluje anomalne punkty w zbiorze danych. Główną zaletą tego podejścia jest bardzo szybki algorytm o niskim zapotrzebowaniu na pamięć oraz możliwość pracować z dużym zbiorem danych.

## 4. Cel projektu

Celem projektu jest opracowanie i implementacji algorytmu *Isolation Forest* do wykrywania anomalii. Sprawdzenia algorytmu na wybranych zbiorach danych. Porównanie danego algorytmu z istniejącymi algorytmami klasyfikacji.

## 5. Zakres Projektu

Dla osiągnięcia celu (Punkt 4), prace nad projektem można podzielić na kilka części:

- 1) Opracowanie oraz implementacji algorytmu.
- 2) Zbieranie i przetwarzanie zbioru danych. Zbiór danych zostanie wybrany z listy "Outlier Detection DataSets (ODDS)" [4].
- 3) Testy sprawdzające poprawność weryfikacji anomalii dla różnych przypadków (np. parametry wejściowe, ilość danych, typ danych itp.).
- 4) Zaimplementowanie innych algorytmów klasyfikacji dla wykrywania anomalii za pomocą istniejących już bibliotek.
- 5) Porównywanie działania algorytmu z innymi dostępnymi w języku R metodami detekcji wartości odstających.
- 6) Napisanie wniosków.

## 6. Opis algorytmu

Zasada działania algorytmu obejmuje dwie fazy, pierwszą uczącą oraz drugą predykcyjną. Podczas uczenia tworzony jest las losowy składający się z określonej liczby drzew losowych (oryginalna terminologia). Ponieważ metoda izolacyjnego lasu losowego jest nienadzorowanym algorytmem uczenia maszynowego to nie występuje potrzeba oznaczania danych pod kątem odstawiania przed fazą nauki. Podczas treningu tworzony jest las składający się z drzew losowych takich, że każdy węzeł drzewa "T", jest albo liściem albo wewnętrznym węzłem, z dokładnie jednym warunkiem podziału oraz dokładnie dwoma węzłami-córkami. Węzły pochodne "Tl" i "Tr", oznaczane są od tego na którą stronę rozdzielają płaszczyznę w podziale. Węzeł składa się z oznaczenia atrybutu "a" na podstawie którego

przeprowadzany jest test oraz wartość graniczna "p" taka, że porównanie wartości "a" w teście "p" daje jednoznaczną odpowiedź czy punkt leży po prawej czy lewej stronie podziału. Wartość graniczna jest wartością pomiędzy wartością minimalną i maksymalną na zbiorze trenującym dla określonego atrybutu i jest wybierana losowo. Podczas podziału następuje rekursywne tworzenie nowych węzłów aż do momentu gdy w przedziale znajduje się tylko jedna wartość, osiągnięto zakładaną maksymalną głębokość lub wszystkie dane mają tę samą wartość atrybutów. Jako prawidłowo utworzony las uznaje się taki który zawiera liczbę liści równą  $\psi$  (liczba przykładów trenujących), a liczba węzłów wewnętrznych wynosi  $\psi-1$ . W ten sposób stworzone drzewo posiada  $2\psi-1$  węzłów.

W zadaniu detekcji anomalii ważny jest wskaźnik opisujący z jakim prawdopodobieństwem dany pomiar jest uznany za wartość odstającą. W algorytmie izolacyjnego lasu losowego tym wskaźnikiem jest długość drzewa którą trzeba pokonać do wyodrębnienia wartości. Wartości odstające cechują się tym, że do ich wyodrębnienia potrzeba zdecydowanie mniej podziałów niż w przypadku wartości standardowych. Długość ścieżki "h(x)" dla danej wartości "x" jest sumą gałęzi które trzeba pokonać od korzenia do liścia.

Problemem przy określaniu i porównywaniu wskaźnika anomalii "h(x)" jest to, że maksymalna wielkość drzewa rośnie proporcjonalnie do ilości danych treningowych to średnia wielkość rośnie według skali logarytmicznej. Dlatego, żeby możliwe było porównywanie wartości konieczna jest normalizacja co zostało pokazane we [Wzór 1.]. Wysoka wartość wskaźnika anomalii (bliska 1) oznacza, że wartość jest odstająca. Wartości zdecydowanie mniejsze niż 0.5 najprawdopodobniej są wartościami normalnymi.

$$S = 2^{-\left(\frac{\text{Average path length for observation}}{\text{Average length of all path}}\right)}$$

Wzór 1. Wynik anomalii.

## 7. Plan badań

Jako pierwszy etap postanowiono na ręcznie wygenerowanym sztucznym zbiorze, oraz na publicznie dostępnych zbiorach [4] z oznaczeniem wartości odstających sprawdzić działanie algorytmu.

Jako drugi etap założono sprawdzenie wpływu poszczególnych parametrów na jakość uzyskiwanej izolacji wartości odstających w tym szczególnie:

- a) Maksymalną wielkość drzewa w fazie treningu
- b) Wielkość zbioru treningowego
- c) Maksymalną wielkość drzewa w fazie ewaluacji

Trzecim etapem badania będzie porównanie działania algorytmu z innymi dostępnymi w języku R metodami detekcji wartości odstających w tym szczególnie:

- a) K najbliższych sąsiadów (angl. *k-nearest neighbors* - K-NN)
- b) ORCA
- c) Klasyfikacja jednoklasowa (angl. *One-Class SVM*)
- d) Lokalny współczynnik wyjątkowości (angl. *Local outlier factor* - LOF)
- e) Las losowy (angl. *Random Forests*)

Jako ostatni etap zostanie przeprowadzona ocena działania algorytmu. Na obecnym etapie postanowiono użyć metryki AUC. Główna idea to wyliczenie pola pod krzywą (angl. *Area Under ROC Curve*). Ta miara pomoże ocenić skuteczność klasyfikatorów (od 0 do 1): im większy AUC tym lepiej. Czyli w przypadku gdy wynik AUC będzie 1, to klasyfikator doskonale klasyfikuje obserwacje, wynik  $\leq 0.5$  wskazując, że model jest losowy albo gorszy niż losowy [5].

## 8. Otwarte kwestie wymagające późniejszego rozwiązania

1. Na jakich zbiorach powinniśmy testować algorytm?
2. Czy powinniśmy zakładać istnienie danych tekstowych i ich ewentualną zamianę na wartości liczbowe/ pomijanie?
3. Czy wskazane przez nas algorytmy klasyfikacji z którymi będziemy chcieli porównać działanie są dobrze wybrane?
4. Podczas treningu co powinniśmy zrobić gdy warunek podziału jest równy wartości atrybutu?
5. Czy metryka AUC zawsze jest dobrą metodą do oceny działania tego algorytmu i jej rezultat nie może wprowadzić w błąd.

## Źródła danych:

- [1] Liu, Fei Tony, Ting, Kai Ming i Zhou, Zhi-Hua. „Las izolacji”. Data Mining, 2008. ICDM'08. Ósma Międzynarodowa Konferencja IEEE nt.
- [2] Liu, Fei Tony, Kai Ming Ting, Zhi-Hua Zhou. "Isolation-based anomaly detection." ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012): 1-39..
- [3] "Isolation forest" [https://en.wikipedia.org/wiki/Isolation\\_forest](https://en.wikipedia.org/wiki/Isolation_forest) [Online]. Accessed: 12 April 2021.
- [4] "Outlier Detection DataSets (ODDS)" <http://odds.cs.stonybrook.edu/> [Online]. Accessed: 12 April 2021.
- [5] Ryan Gillespie. "Detecting Fraud and Other Anomalies Using Isolation Forest". SAS Institute Inc., Cary, NC.