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.

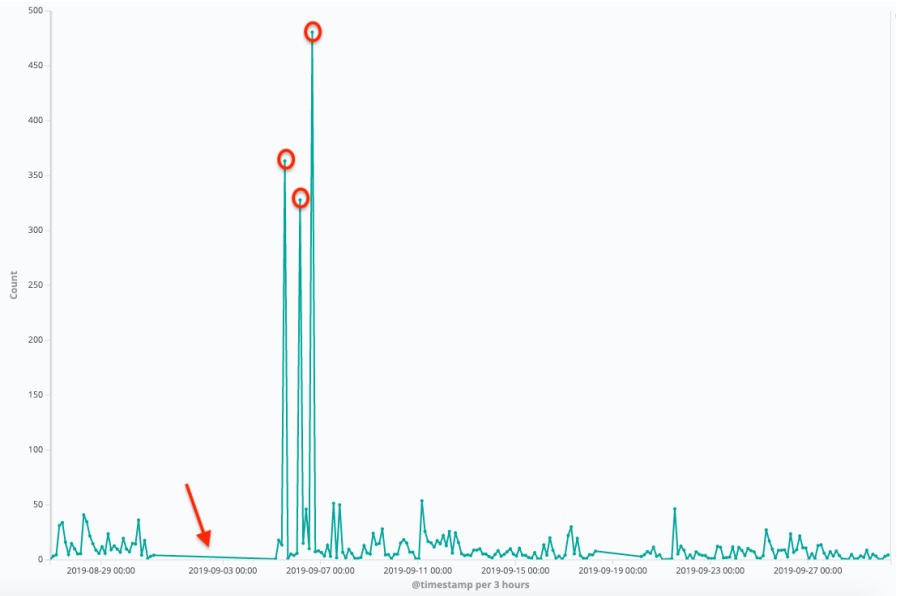
2. Skład zespołu

- 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 algorytmu zwykłe 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 styczność 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 wybranych zbiorów danych.
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 odstawania 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. Atrybuty nie ciągłe są porównywane równościowo. 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 tą samą wartości atrybutów. Jako prawidłowo utworzony las (bez uwzględniania warunków na maksymalną wysokość drzewa) uznaje się taki który zawiera liczbę liści równą ψ (liczba przykładów trenujących), a liczba węzłów wewnętrznych wynosi ψ-1. W ten sposób stworzone drzewo posiada 2ψ-1 węzłów. Dla modelu o ograniczonym rozmiarze te parametry mogą nie spełniać wyżej podanej zależności.

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.

$$S = 2^{-\left(\frac{Average\ path\ length\ for\ observation}{Average\ path\ length\ of\ unsuccessful\ searches\ in\ BST}\right)}$$

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

Testy postanowiono przeprowadzić na zbiorach stosowanych do detekcji:

a) Credit Card Fraud Detection (Kaggle) - Wybrany ze względu bardzo dobrze udokumentowany zbiór oraz wykorzystanie częste wykorzystanie go do pokazania możliwości algorytmu Isolation Forest.

b) Arrhythmia dataset (ODDS) - Wybrany ze względu na bardzo dużą liczbę klas w stosunku do ilości przykładów oraz dużą liczbę wartości odstających (~15%).
c) http (KDDCUP99) dataset (ODDS) - Wybrany ze względu na posiadanie wartości nienumerycznych oraz bardzo mały współczynnik wartości odstających oraz małą liczbę klas.

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*) - Algorytm nadzorowany, którego działanie polega na wyborze wartości dla przykładu na podstawie k (parametr) liczby najbliższych mu wartości według przyjętej metryki i przypisaniu wartości na podstawie wartości sąsiadów.
b) Klasyfikacja jednoklasowa (angl. *One-Class SVM*) (OCC) - polega na zidentyfikowaniu najmniejszej hipersfery (o promieniu r i środku c) składającej się ze wszystkich punktów danych. Uczenie jest odmianą uczenia PU , w którym klasyfikator binarny jest uczony w sposób częściowo nadzorowany tylko z pozytywnych i nieoznakowanych punktów próbki. W nauczaniu PU zakłada się, że do uczenia dostępne są dwa zestawy przykładów: zbiór pozytywny i zestaw mieszany , który zakłada, że zawiera zarówno próbki pozytywne, jak i negatywne, ale bez oznaczania ich jako takich. (R - paczka One-class)
c) Lokalny współczynnik wyjątkowości (angl. *Local outlier factor - LOF*) - Lokalny współczynnik wartości odstających opiera się na koncepcji gęstości lokalnej, w której lokalność określa k najbliższych sąsiadów, których odległość służy do szacowania gęstości. Uczenie algorytmu jest nienadzorowane. Porównując lokalną gęstość obiektu z lokalną gęstością jego sąsiadów, można zidentyfikować regiony o podobnej gęstości oraz punkty, które mają znacznie mniejszą gęstość niż ich sąsiedzi. Są one uważane za wartości odstające. (R - DDoutlier)
d) Las losowy (angl. *Random Forests*) - Algorytm nadzorowanego uczenia maszynowego polegający na zgrupowaniu wielu binarnych drzew decyzyjnych których odpowiedzi są łączone i dają jeden wynik zwykle większościowy (R - randomForest)

Jako ostatni etap zostanie przeprowadzona ocena działania algorytmu. Na obecnym etapie postanowiono użyć metryki precision-recall. Precyzja jest stosunkiem liczby prawdziwych pozytywnych przykładów podzielonych przez sumę prawdziwych i fałszywych pozytywów. Opisuje ona, jak dobry jest model w przewidywaniu klasy pozytywnej. Precyzja jest określana jako pozytywna wartość predykcyjna. Recall (czułość) jest obliczana jako stosunek liczby wyników prawdziwie pozytywnych do sumy wyników prawdziwie pozytywnych i fałszywie negatywnych. Ta miara pomoże ocenić skuteczność klasyfikatorów, przeglądanie zarówno precyzji, jak i czułości jest przydatne w przypadkach, gdy występuje nierównowaga w obserwacjach między dwiema klasami. Konkretnie, istnieje wiele przykładów braku zdarzenia (klasa 0) i tylko kilka przykładów zdarzenia (klasa 1).

8. Otwarte kwestie wymagające późniejszego rozwiązania

1. Czy wybrane zbiory danych są odpowiednie?
2. Czy ze względu na trudności w znalezieniu odpowiedniej paczki z implementacją algorytmu ORCA w języku R możemy zrezygnować z porównania działania algorytmu z naszą implementacją?
3. Czy powinniśmy zakładać istnienie danych tekstowych i ich ewentualną zamianę na wartości liczbowe/ pomijanie? - odpowiedź jest jasna
4. Czy wskazane przez nas algorytmy klasyfikacji z którymi będziemy chcieli porównać działanie są dobrze wybrane?- odpowiedź jest jasna
5. Czy metryka AUC zawsze jest dobrą metodą do oceny działania tego algorytmu i jej rezultat nie może wprowadzić w błąd. - odpowiedź jest jasna

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 [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.