Wykrywanie anomalii w logach dostępu do serwera Apache za pomocą algorytmu KNN

Na podstawie:
„Detecting anomalous Web server usage through mining access logs” T. Gržinić, T. Kišasondi, J. Šaban

Adam Chyła

Wykrywanie anomalii w logach dostępu do serwera Apache za pomocą algorytmu KNN

Thanks to Toni Gržinić for answering my questions about article.

Adam Chyła

Agenda

- Metoda rozstępu międzykwartylowego (metoda IQR)
- Przygotowanie danych do analizy
  - Wyodrębnienie poszczególnych sesji użytkowników
  - Obliczenie statystyk dla poszczególnych sesji
  - Oznaczenie sesji jako anomalii za pomocą metody IQR
- Analiza danych za pomocą algorytmu KNN
- Wyniki osiągnięte przez autorów artykułu

Agenda

- **Metoda rozstępu międzykwartylowego (metoda IQR)**
- Przygotowanie danych do analizy
  - Wyodrębnienie poszczególnych sesji użytkowników
  - Obliczenie statystyk dla poszczególnych sesji
  - Oznaczenie sesji jako anomalii za pomocą metody IQR
- Analiza danych za pomocą algorytmu KNN
- Wyniki osiągnięte przez autorów artykułu

Metoda rozstępu międzykwartylowego (metoda IQR)

Metoda używana do określenia obserwacji odstających (anomalii) z pośród podanego posortowanego niemalejąco zbioru liczb.

Polega na:

- obliczeniu różnicy (IQR) pomiędzy kwartylem trzecim (Q3) a kwartylem pierwszym (Q1)
- oznaczeniu wszystkich wartości mniejszych od Q1 − 1.5 ∗ IQR jako wartości odstających
- oznaczeniu wszystkich wartości większych od Q3 + 1.5 ∗ IQR jako wartości odstających

Metoda rozstępu międzykwartylowego (metoda IQR)

Przykład:

Dany jest zbiór:

M = {3, 6, 6.5, **7**, 8, 8.5, 9, 9, 9, 9.5, 10, **10**, 11, 12, 19}

Wyznaczamy Q1, Q3, IQR:

Q1 = 7 (gdyż jest to 4 liczba w zbiorze -> |M| / 4 =15 / 4 = 3.75)

Q3 = 10 (gdyż jest to 12 liczba w zbiorze -> 3 * |M| / 4 = 11.25)

IQR = Q3 - Q1 = 3

Metoda rozstępu międzykwartylowego (metoda IQR)

Przykład:

Dany jest zbiór:

M = {3, 6, 6.5, **7**, 8, 8.5, 9, 9, 9, 9.5, 10, **10**, 11, 12, 19}

Q1 = 7; **Q3** = 10; **IQR** = 3

Oznaczamy wartości odstające:

a) mniejsze od Q1 − 1.5 ∗ IQR = 7 − 1.5 ∗ 3 = **2.5** *(brak takich wartości w zbiorze)*

b) Większych od Q3 + 1.5 ∗ IQR = 10 + 1.5 ∗ 3 = **14.5** (liczba 19)

Agenda

- Metoda rozstępu międzykwartylowego (metoda IQR)
- **Przygotowanie danych do analizy**
  - Wyodrębnienie poszczególnych sesji użytkowników
  - Obliczenie statystyk dla poszczególnych sesji
  - Oznaczenie sesji jako anomalii za pomocą metody IQR
- Analiza danych za pomocą algorytmu KNN
- Wyniki osiągnięte przez autorów artykułu

Przygotowanie danych do analizy

Wykonania analizy logów dostępu do serwera Apache jest możliwe, gdy są one zapisane w formacie **Combined Log**.

W odróżnieniu od **Common Log Format** (domyślnie stosowany przez Apache) zawiera on dodatkowe pola:

- referer – adres URL poprzednio odwiedzonej przez użytkownika strony
- user-agent – informacja o kliencie użytkownika

Przygotowanie danych do analizy

Do wykonania analizy logów dostępu do serwera Apache zostaną wykorzystane następujące pola:

- adres IP
- znacznik czasu
- zwrócony kod (status code)
- ilość wysłanych danych
- identyfikator klienta (user-agent)

Przygotowanie danych do analizy

- Plik z logami należy podzielić na części.
- Każda z części to jeden dzień działania serwera Apache (24 godziny).
- W każdej części dane przygotowywane są niezależnie.

Agenda

- Metoda rozstępu międzykwartylowego (metoda IQR)
- Przygotowanie danych do analizy
  - **Wyodrębnienie poszczególnych sesji użytkowników**
  - Obliczenie statystyk dla poszczególnych sesji
  - Oznaczenie sesji jako anomalii za pomocą metody IQR
- Analiza danych za pomocą algorytmu KNN
- Wyniki osiągnięte przez autorów artykułu

Wyodrębnienie poszczególnych sesji użytkowników

Na podstawie wpisów (reprezentujących zapytania) w logach należy wyróżnić poszczególne sesje użytkownika.

Do wykonania tego zadania zostaną wykorzystane pola:

- znacznik czasu
- adres IP
- Identyfikator przeglądarki (user-agent)

Wyodrębnienie poszczególnych sesji użytkowników

Założenie: sesja użytkownika nie trwa dłużej, niż godzinę od czasu wystąpienia pierwszego zapytania.

W przypadku przekroczenia tego czasu należy uznać, iż użytkownik rozpoczął nową sesję.

Jeśli zapytania zawierają ten sam adres IP oraz identyfikator klienta (user-agent) należą one do tej samej sesji.

Wyodrębnienie poszczególnych sesji użytkowników

Przykład:

```
znacznik czasu     adres IP  id klienta    wytypowana sesja
2015-9-1:11:23:22  10.0.1.4  UA1 --------> S1
2015-9-1:11:23:12  10.0.1.4  UA1 --------> S1
2015-9-1:11:23:32  10.0.1.4  UA1 --------> S1
...
2015-9-1:11:26:02  10.0.1.5  UA3 --------> S2
2015-9-1:11:27:02  10.0.1.5  UA3 --------> S2
...
2015-9-1:13:23:32  10.0.1.4  UA1 --------> S3
...
```

* wpisy z logów zostały zapisane w zmienionym formacie, by poprawić czytelność

Agenda

- Metoda rozstępu międzykwartylowego (metoda IQR)
- Przygotowanie danych do analizy
  - Wyodrębnienie poszczególnych sesji użytkowników
  - **Obliczenie statystyk dla poszczególnych sesji**
  - Oznaczenie sesji jako anomalii za pomocą metody IQR
- Analiza danych za pomocą algorytmu KNN
- Wyniki osiągnięte przez autorów artykułu

Obliczenie statystyk dla poszczególnych sesji

Dla każdej sesji należy obliczyć statystyki zawierające:

- czas trwania sesji (różnica czasu pomiędzy ostatnim a pierwszym zapytaniem w sesji w sekundach)
- użycie pasma sieciowego (ilości pobranych danych w bajtach)
- całkowita liczba zapytań w danej sesji
- procent błędnych zapytań (kod błędu jest pomiędzy 400 a 500)

Obliczenie statystyk dla poszczególnych sesji

Przykład:

sesja	czas trwania	zużycie pasma	l. zapytań	% błędnych
S1	10s	11034b	56	33
S2	0s	1043b	2	0
S3	60s	12004b	5	50
S4	0s	90b	12	0
S5	15s	91b	5	0
S6	7s	30b	8	0
S7	19s	197b	112	0
S8	12s	101b	37	0

Agenda

- Metoda rozstępu międzykwartylowego (metoda IQR)
- Przygotowanie danych do analizy
  - Wyodrębnienie poszczególnych sesji użytkowników
  - Obliczenie statystyk dla poszczególnych sesji
  - **Oznaczenie sesji jako anomalii za pomocą metody IQR**
- Analiza danych za pomocą algorytmu KNN
- Wyniki osiągnięte przez autorów artykułu

Oznaczenie sesji jako anomalii za pomocą metody IQR

Oznaczenie sesji jako anomalii następuje na podstawie wykonanej ilości zapytań.

Zapisujemy ilość zapytań z każdej sesji w porządku niemalejącym:

S2	S3	S5	S6	S4	S8	S1	S7
2	5	5	8	12	37	56	112

Oznaczenie sesji jako anomalii za pomocą metody IQR

Oznaczenie sesji jako anomalii następuje na podstawie wykonanej ilości zapytań.

S2	**S3**	S5	S6	S4	**S8**	S1	**S7**
2	**5**	5	8	12	**37**	56	**112**

Q1 = 5; **Q3** = 37; **IQR** = 32

Oznaczamy jako anomalie sesje, których ilość zapytań jest:

- mniejsza od Q1 − 1.5 ∗ IQR = −43 (nie możliwe)
- większa od Q3 + 1.5 ∗ IQR = 37 + 48 = 85 (tylko sesja **S7**)

Oznaczenie sesji jako anomalii za pomocą metody IQR

Tak oznaczone sesje należy przejrzeć i ewentualnie poprawić ich klasyfikację. Oznaczanie sesji jako anomalii metodą IQR jest tylko pewnym przybliżeniem.

Do uzyskania lepszej jakości klasyfikacji wpisów jako anomalii można dodatkowo użyć reguł – np. pochodzących z systemów wykrywania włamań (IDS).

Agenda

- Metoda rozstępu międzykwartylowego (metoda IQR)
- Przygotowanie danych do analizy
  - Wyodrębnienie poszczególnych sesji użytkowników
  - Obliczenie statystyk dla poszczególnych sesji
  - Oznaczenie sesji jako anomalii za pomocą metody IQR
- **Analiza danych za pomocą algorytmu KNN**
- Wyniki osiągnięte przez autorów artykułu

Analiza danych za pomocą algorytmu KNN

Algorytm K-najbliższych sąsiadów (*K-nearest neighbours*) służy do kategoryzacji nowych obserwacji na podstawie znanych obserwacji i ich kategorii.

Analiza danych za pomocą algorytmu KNN

Algorytm przyjmuje:

[Z] – lista wektorów cech znanych obserwacji

K – wektor znanych kategorii odpowiadających obserwacjom

N – wektor cech nowej obserwacji

k – liczba sąsiadów, na podstawie których określić klasyfikację nowej obserwacji

Algorytm zwraca liczbę odpowiadającą kategorii danej obserwacji.

Analiza danych za pomocą algorytmu KNN

Przykład danych wejściowych:

```
Lista wektorów znanych sesji     Wektor K znanych kategorii:
(ze wszystkich części) [Z]:

[[10, 11034, 56, 33],             [0,
 [0,  1043,  2,  0],               0,
 [60, 12004, 5,  50],              0,
 [0,  90,    12, 0],               0,
 [15, 91,    5,  0],               0,
 [7,  30,    8,  0],               0,
 [19, 197, 112,  0],               1,
 [12, 101,  37,  0]]               0]
```

Umowa: kategoria 0 oznacza brak anomalii, 1 oznacza anomalię

Analiza danych za pomocą algorytmu KNN

Przykład danych wejściowych:

```
Wektor cech nowej obserwacji:    Liczba k sąsiadów:
[18, 190, 113,  0]               1
```

Przykład danych wyjściowych:

Liczba odpowiadająca danej kategorii. W tym przykładzie odpowiedzią będzie 1 – kategoria oznaczająca anomalię.

Agenda

- Metoda rozstępu międzykwartylowego (metoda IQR)
- Przygotowanie danych do analizy
  - Wyodrębnienie poszczególnych sesji użytkowników
  - Obliczenie statystyk dla poszczególnych sesji
  - Oznaczenie sesji jako anomalii za pomocą metody IQR
- Analiza danych za pomocą algorytmu KNN
- **Wyniki osiągnięte przez autorów artykułu**

Wyniki osiągnięte przez autorów artykułu

Zastosowano przedstawioną metodę analizy danych z drobnymi różnicami.

Autorzy po przygotowaniu danych do analizy podzielili dane na dwie części. Pierwsza część to **25%** wszystkich danych, służyła jako znane obserwacje dla algorytmu KNN. Druga część służyła ocenie jakości klasyfikacji.

Algorytm zaklasyfikował poprawnie **60,60%** obserwacji z drugiej części zbioru.

Wyniki osiągnięte przez autorów artykułu

Zastosowano przedstawioną metodę analizy danych z drobnymi różnicami.

Autorzy po przygotowaniu danych do analizy podzielili dane na dwie części. Pierwsza część to **50%** wszystkich danych, służyła jako znane obserwacje dla algorytmu KNN. Druga część służyła ocenie jakości klasyfikacji.

Algorytm zaklasyfikował poprawnie **86.27%** obserwacji z drugiej części zbioru.

Dziękuję za uwagę.