Przetwarzanie i analiza dużych zbiorów danych2021/2022

Prowadzący: mgr inż. Rafał Woźniak Czwartek, 15:45

Szymon Gruda	239661	239661@edu.p.lodz.pl
Jan Karwowski	239671	239671@edu.p.lodz.pl
Michał Kidawa	239673	239673@edu.p.lodz.pl
Kamil Kowalewski	239676	239676@edu.p.lodz.pl

Checkpoint 2

Spis treści

1. Wprowadzenie

Głównym celem projektu jest przeprowadzenie kompleksowej analizy zbioru *NYPD Complaint Data Historic* [1]. Jego dokładny opis został wykonany w ramach Checkpoint 1 więc w sekcji 3 jest on przedstawiony w dużo bardziej zwięzły sposób. Cele projektu pod dokonaniu modyfikacji zgodnie z uwagami z Checkpoint 1 zostały przedstawione w sekcji 4.

2. Podział obowiązków w zespole

— Szymon Gruda - Zbadanie możliwości realizacji celu nr 1 i analiza wyników.
— Jan Karwowski - Wstępne podejście do realizacji celu nr 2
— Michał Kidawa - Wstępne podejście do realizacji celu nr 3
— Kamil Kowalewski - Wykonanie pełnego preprocessingu danych łącznie z jego opisem oraz przeniesienie i dostosowanie treści sprawozdania z Checkpoint 1 do Checkpoint 2

3. Charakterystyka zbioru danych

Kolumny w zbiorze danych zdecydowaliśmy sie podzielić na następujące grupy:
1. Identyfikator
2. Data i czas zdarzenia
3. Data i czas zgłoszenia
4. Typ i opis wykroczenia/przestępstwa
5. Czy się udało
6. Otoczenie zdarzenia
7. Lokalizacja zdarzenia
8. Cechy podejrzanego
9. Cechy ofiary

4. Cele projektu

W ramach projektu sformułowane zostały trzy następujące cele.

4.1. Klasyfikacja rodzaju lub poziomu przestępstwa

W ramach tego etapu przeprowadzony zostanie szereg eksperymentów mających na celu stworzenie klasyfikatora typu przestępstwa (KY_CD). Jeżeli okaże się to niemożliwe podjęta zostanie próba klasyfikacji poziomu wykroczenia, stanowiącego bardziej ogólną informacje. Klasyfikacja ta będzie odbywała się na podstawie następujących informacji:
— godzina zdarzenia
— dzień tygodnia zdarzenia
— odstęp między zgłoszeniem a zdarzeniem
— czy doszło do skutku (CRM_ATPT_CPTD_CD)

— otoczenie zdarzenia
— cechy podejrzanego
— cechy ofiary
— poziom lub typ przestępstwa w zależności od tego co klasyfikujemy

Wybrane cechy mogą ulec drobnym modyfikacjom w trakcie trwania eksperymentów. Opcjonalnie podjęta będzie próba wykorzystania informacji o czasie trwania przestępstwa (dla tych zdarzeń, dla których jest dostępna).

Informacja o dokładnej lokalizacji zdarzenia nie jest wykorzystana ze względu na chęć stworzenia uniwersalnego narzędzia.

Aby zrealizować zaproponowany cel wykorzystane zostaną następujące metody: metody imputacji brakujących danych, prosta ekstrakcja cech (zwłaszcza z daty), naiwny klasyfikator Bayesa i lasy losowe.

4.2. Regresja czasu trwania przestępstwa

Drugim celem będzie regresja czasu trwania przestępstwa dla tych przestępstw, dla których jest on znany. Tak więc wybrany zostanie podzbiór głównego zbioru danych i na podstawie czasu rozpoczęcia i zakończenia przestępstwa wyekstrahowany zostanie czas jego trwania. Na podstawie pozostałych kolumn (poza tych z datą i czasem) podjęta będzie próba wyuczenia modelu regresyjnego estymowania tego czasu.

Wykorzystane zostaną modele przystosowane do zadań regresji, tak więc przede wszystkim lasy losowe, ewentualnie klasyfikator liniowy, maszyny wektorów nośnych czy wreszcie sieć neuronowa. Wykorzystane metody będą dobierane odpowiednio w zależności od wyników otrzymywanych w trakcie trwania eksperymentów i od napotkanych problemów. Dodatkowo planowane jest wykorzystanie biblioteki *XGBoost*.

4.3. Grupowanie przestępstw na podstawie podzbiorów cech

W ramach tego celu przeprowadzone zostanie automatyczne grupowanie zdarzeń. Spośród cech podejrzanego i ofiary kilkakrotnie wybrany zostanie podzbiorów stanowiących przedmiot eksperymentu - etykietę. Dla każdej wybranej etykiety (która może być złożeniem kilku atrybutów) przeprowadzona zostanie seria eksperymentów, mająca na celu automatyczne pogrupowanie zdarzeń zgodnie z tą etykietą, wykorzystując wszystkie pozostałe atrybuty (poza tworzącymi etykietę). Po przeprowadzeniu grupowania z wykorzystaniem kilku różnych algorytmów (DBSCAN, k-means, algorytm aglomeracyjny), zmierzona zostanie jakość grupowania za pomocą metryk zewnętrznych (accuracy) względem wybranej etykiety. Dodatkowo porównana zostanie jakość grupowania między seriami eksperymentów (dla różnych etykiet) za pomocą metryk wewnętrznych.

5. Wstępne przetwarzanie danych

W ramach wstępnego przetwarzania danych został przygotowany skrypt o nazwie *dataset_preprocessing.py*. Miał on do wykonani kilka zadań, pierwszym z nich było połączenie następujących kolumn *CMPLNT_FR_DT* i

CMPLNT_FR_TM oraz *CMPLNT_TO_DT* i *CMPLNT_TO_TM* celem stworzenia nowej kolumny zawierający obiekt *Timestamp*. Co więcej dla łatwiejsze wykorzystania *RPT_DT* również został przekonwertowany na obiekt *Timestamp*. Drugim zadaniem było wygenerowanie słownika kodów przestępstw, było to dokonane dla kolumn *KY_CD* oraz *PD_CD* celem sprawdzenia czy każdy kod odpowiada temu samemu opisowi lub znaczenia opisu są tożsame. Trzecim zadaniem było uporządkowanie kolumn dotyczących rasy oraz płci sprawcy i ofiary. W przypadku rasy dla wartości *nan*, *UNKNOWN*, *OTHER* została ustawiona wartość *OTHER*. Ze względu na specyficzne dane znajdujące się w kolumnach określających płeć zdecydowaliśmy się na zmianę wartości z *F* na *FEMALE*, z *M* na *MALE* a w pozostałych przypadkach była wstawiana wartość *OTHER*. Czwartym zadaniem było usuniecię nadmiarowych kolumn lub tych na podstawie, których zostały wygenerowane nowe. Przed ostatnim zadaniem było wygenerowanie statystyk o tym ile brakuje danym w wybranych kolumnach. Ostatnim zadaniem był zapis zmodyfikowanego zbioru danych do pliku *.csv*.

6. Przetwarzanie i analiza danych

6.1. Klasyfikacja rodzaju lub poziomu przestępstwa

W celu realizacji tego zadania, z daty wystąpienia przestępstwa wyekstrahowane zostały informacje dotyczące dnia tygodnia oraz godziny wystąpienia zdarzenia. Na tym etapie badań, wszelkie wiersze, które posiadały brakujące dane zostały usunięte. Wykorzystując fragment danych, przeprowadzone zostały badania mające na celu ustalić najodpowiedniejsze parametry klasyfikatorów (las losowy i naiwny klasyfikator Bayes'a)

6.1.1. Wyniki

W tabelach 1 i 2 przedstawione zostały wartości parametrów (również procent podziału danych na zbiór uczący i testowy) i ich wpływ na dokładność klasyfikacji dla 400000 wierszy, z których 73992 zostały odrzucone na skutek brakujących danych, dla klasyfikatora lasów losowych. Tabela **??** przedstawia wyniki otrzymane dla różnego procentowego podziału danych na zbiór uczący i testowy.

% testowy	Nazwa parametru	Wartość parametru	Skuteczność trening	skuteczność test
10%	min_samples_leaf	10	0.6904	0.6161
10%	min_samples_leaf	100	0.6925	0.6288
10%	min_samples_leaf	1000	0.6146	0.6063
10%	min_samples_leaf	10000	0.5787	0.5702
20%	min_samples_leaf	10	0.6893	0.6199
20%	min_samples_leaf	100	0.6398	0.6353
20%	min_samples_leaf	1000	0.6105	0.6160
20%	min_samples_leaf	10000	0.5737	0.5780
30%	min_samples_leaf	10	0.6894	0.6152
30%	min_samples_leaf	100	0.6405	0.6314
30%	min_samples_leaf	1000	0.6125	0.6086
30%	min_samples_leaf	10000	0.5701	0.5668
35%	min_samples_leaf	10	0.6904	0.6155
35%	min_samples_leaf	100	0.6411	0.6300
35%	min_samples_leaf	1000	0.6127	0.6083
35%	min_samples_leaf	10000	0.5630	0.5610
10%	max_depth	100	0.9255	0.5450
10%	max_depth	1000	0.9255	0.5450
10%	max_depth	10000	0.9255	0.5450
10%	max_depth	100000	0.9255	0.5450
20%	max_depth	100	0.9293	0.5533
20%	max_depth	1000	0.9293	0.5533
20%	max_depth	10000	0.9293	0.5533
20%	max_depth	100000	0.9293	0.5533
30%	max_depth	100	0.9348	0.5474
30%	max_depth	1000	0.9348	0.5474
30%	max_depth	10000	0.9348	0.5474
30%	max_depth	100000	0.9348	0.5474
35%	max_depth	100	0.9372	0.5474
35%	max_depth	1000	0.9372	0.5474
35%	max_depth	10000	0.9372	0.5474
35%	max_depth	100000	0.9372	0.5474
10%	n_estimators	100	0.6342	0.6235
10%	n_estimators	1000	0.6342	0.6236
20%	n_estimators	100	0.6305	0.6338
20%	n_estimators	1000	0.6310	0.6348
30%	n_estimators	100	0.6310	0.6275
30%	n_estimators	1000	0.6312	0.6274
35%	n_estimators	100	0.6321	0.6257
35%	n_estimators	1000	0.6321	0.6254

Tabela 1. Wyniki dla poszczególnych parametrów dla klasyfikatora lasów losowych

% testowy	Nazwa parametru	Wartość parametru	Skuteczność trening	skuteczność test
10%	min_samples_leaf	10	0.6904	0.6161
10%	max_samples	0.01	0.5667	0.5548
10%	max_samples	0.02	0.5859	0.5750
10%	max_samples	0.05	0.6055	0.5947
10%	max_samples	0.1	0.6148	0.6049
10%	max_samples	0.2	0.6213	0.6115
10%	max_samples	0.5	0.6291	0.6192
10%	max_samples	0.8	0.6445	0.6221
10%	max_samples	0.99	0.6339	0.6232
20%	max_samples	0.01	0.5609	0.5637
20%	max_samples	0.02	0.5826	0.5875
20%	max_samples	0.05	0.6021	0.6075
20%	max_samples	0.1	0.6118	0.6170
20%	max_samples	0.2	0.6187	0.6233
20%	max_samples	0.5	0.6258	0.6303
20%	max_samples	0.8	0.6293	0.6332
20%	max_samples	0.99	0.6308	0.6348
30%	max_samples	0.01	0.5446	0.5448
30%	max_samples	0.02	0.5822	0.5815
30%	max_samples	0.05	0.6034	0.6007
30%	max_samples	0.1	0.6122	0.6097
30%	max_samples	0.2	0.6184	0.6156
30%	max_samples	0.5	0.6266	0.6236
30%	max_samples	0.8	0.6295	0.6258
30%	max_samples	0.99	0.6313	0.6279
35%	max_samples	0.01	0.5435	0.5424
35%	max_samples	0.02	0.5736	0.5718
35%	max_samples	0.05	0.6024	0.5996
35%	max_samples	0.1	0.6114	0.6074
35%	max_samples	0.2	0.6190	0.6141
35%	max_samples	0.5	0.6272	0.6213
35%	max_samples	0.8	0.6305	0.6237
35%	max_samples	0.99	0.6317	0.6252

Tabela 2. Wyniki dla poszczególnych parametrów dla klasyfikatora lasów losowych

% testowy	Skuteczność
10%	0.328
20%	0.3368
30%	0.3293
35%	0.3284

Tabela 3. Wyniki dla naiwnego klasyfikatora Bayes'a

6.1.2. Analiza wyników

W oparciu o wyniki zamieszczone w tabelach 1 i 2 dla klasyfikatora lasów losowych wybrano wartości parametrów, które umożliwią najwyższą dokładność klasyfikacji i umieszczono je w tabeli 4.

% testowy	Nazwa parametru	Wartość parametru
20%	n_estimators	1000
20%	min_samples_leaf	100
20%	max_samples	0.99
20%	max_depth	100

Tabela 4. Wyniki dla naiwnego klasyfikatora Bayes'a

Na podstawie wyników zamieszczonych w tabeli 3 można stwierdzić, że naiwny klasyfikator bayesa ze względu na niską dokładność nie nadaje się do realizacji naszego zadania.

Bazując na parametrach zamieszczonych w tabeli 4 uruchomiono klasyfikator dla większej liczby danych, wyniki czego zostały zaprezentowane w tabeli 5.

Liczba wierszy	Liczba wierszy po odrzuceniu	Dokładność - trening	Dokładność - test
400000	326008	0.6325	0.6294
600000	485043	0.6181	0.6173
700000	563763	0.6125	0.6114

Tabela 5. Wyniki kolejnych uruchomień klasyfikatora lasów losowych

W wynikach z tabeli 5 można zauważyć spadek dokładności klasyfikacji wraz z wzrostem liczby danych. Dlatego najpewniej należy odrzucić próbę klasyfikacji typu przestępstwa, na rzecz bardziej ogólnego poziomu wykroczenia, dla którego możliwe będzie osiągnięcie wyższej dokładności klasyfikacji.

6.2. Regresja czasu trwania przestępstwa

6.2.1. Przygotowanie danych do analizy

Zgodnie z założonym celem z danych wyekstrahowany został czas trwania przestępstwa (w sekundach), przy tym odrzucone zostały wszystkie rekordy, dla których czas ten nie jest znany i co więcej, w ogóle odrzucone zostały wszystkie rekordy, dla których brakuje jakiegokolwiek atrybutu. Wszystkie

max_depth	MAE treningowy	MAE testowy
4	$1544898s \approx 429h$	$1568598s \approx 435h$
5	$1531379s \approx 425h$	$1554616s \approx 431h$
6	$1503298s \approx 417h$	$1530283s \approx 425h$
7	$1482720s \approx 411h$	$1512315s \approx 420h$
10	$1425666s \approx 396h$	$1480827s \approx 411h$
20	$1070366s \approx 297h$	$1528108s \approx 424h$

Tabela 6. Wyniki dla lasu losowego (100 drzew) przy różnych maksymalnych głębokościach

wartości tekstowe zostały zamienione na wartości liczbowe za pomocą klasy LabelEncoder. Ostatecznie ze zbioru po opisanym wcześniej ogólnym wstępnym przetwarzaniu danych, zawierającym ponad 7 mln. rekordów, zostało około 4.5 mln. rekordów. Ze względu na to, że pierwsze eksperymenty będą wykorzystane z użyciem lasu losowego, żadna normalizacja ani standaryzacja nie jest potrzebna.

6.2.2. Eksperymenty

Przeprowadzonych zostało kilka wstępnych eksperymentów, mających na celu wykorzystanie lasu losowego do regresji czasu trwania zdarzenia. We wszystkich 20% zbioru jest wykorzystane na zbiór testowy, pozostała część stanowi zbiór treningowy. Jako miarę jakości regresji wykorzystano *Mean Absolute Error* czyli średnią wartość bezwzględną błędu - różnicy między wartością oczekiwaną a zwróconą przez model. We wszystkich eksperymentach las modelem regresyjnym jest las losowy składający się ze 100 drzew, gdzie ich głębokość jest ograniczana w różnym stopniu. Wyniki zaprezentowane są w tabeli 6.

6.2.3. Dyskusja i wnioski

Przed przystąpieniem do oceny wyników działania klasyfikatora należy zbadać rozkład czasu trwania przestępstwa w całym zbiorze danych. Poniższe statystyki są przybliżone, jako że dokładne wartości nie mają w tym przypadku dużego znaczenia:
— średnia = 248h
— odchylenie standardowe = 4769h (prawie 200 dni)
— minimum = -122h (błąd - należy dodatkowo oczyścić dane)
— maksimum = ponad 10 lat
— mediana = 1380s
— centyl 80 = 11h
— centyl 90 = 40h
Po pierwsze należy wspomnieć, że rozkład tej wartości jest bardzo nierównomierny. Zdecydowana większość zdarzeń ma czas trwania poniżej 10 godzin, a nawet poniżej jednej godziny. Są jedna również przypadki, kiedy zdarzenie trwa w latach, jest ich jednak bardzo mało. Tak więc zbiór ten jest silnie niezbalansowany, jeżeli chodzi o to konkretne zagadnienie. Z tego powodu można spodziewać się znacznych trudności w uczeniu modelu regresyjnego.

Obserwując wyniki z tabeli 6 można stwierdzić, że optymalna głębokość drzewa znajduje się pomiędzy 10 a 20, jako że, w eksperymentach poprzedzających głębokość 10, błąd maleje. Z drugiej strony dla głębokości 20 model jest przetrenowany - błąd zaczął wzrastać na zbiorze testowym. Jeżeli chodzi o samą wartość błędu, to oscyluje ona w okolicach 400 godzin. Biorąc pod uwagę fakt, że zakres wartości czasu trwania przestępstwa to zakres od kilku minut czy sekund do kilku lat, wynik ten można uznać za niosący pewną wartość. Uwzględniając jednak, że zdecydowana większość zdarzeń trwa kilka godzin lub kilkadziesiąt minut, osiągnięta dokładność absolutnie nie jest zadowalająca.

W dalszych badaniach należy zacząć od oczyszczenia zbioru i ponadto wykonania na nim dodatkowego wstępnego przetwarzania, mającego na celu na przykład zrównoważenie go. Należy również spróbować usprawnić model lasu losowego, być może wykorzystując bibliotekę *XGBoost*. Ponadto można skorzystać z innych modeli oraz wprowadzić dodatkową ekstrakcję nowych cech z obecnie wykorzystywanych atrybutów.

6.3. Grupowanie przestępstw na podstawie podzbiorów cech

Do realizacji tej części zadania dane zostały poddane wstępnej obróbce, a następnie usunięto wiersze posiadające puste dane. Grupowanie odbywało się za pomocą 3 różnych algorytmów klasteryzacji (KMeans, DBSCAN, Aglomercyjny). Spośród cech podejrzanego i ofiary zostały wybrane podzbiory stanowiące przedmiot eksperymentu - etykiety. Etykiety mogły składać się z danych pochodzących z jednej, lub wielu kolumn. Można było wyróżnić 4 eksperymenty na podstawie etykiet:

— SUSP_SEX - grupowanie na podstawie płci podejrzanego
— VIC_RACE - grupowanie na podstawie rasy ofiary
— SUSP_SEX_VIC_SEX - tworzona własna etykieta, która ma określać jakiego rodzaju przestępstwo miało miejce pod kątem płci
— SUSP_RACE_VIC_RACE - stworzona własna etykieta, która określa tło rasowe incydentu

6.3.1. Metodologia badań

Na początku każdego eksperymentu wybierana została etykieta, która stanowiła główny przedmiot eksperymentu. Usuwane zostały niepełne, rekordy, a następnie pozostałe etykiety były zamieniane na wartości liczbowe do analizy. W przypadku algorytmów KMeans oraz Aglomeracyjnego liczba klastrów ustawiana była na liczbę unikatowych wartości w ramach danej etykiety. Ponieważ ten fragment zadania przewidywał wstępne przetwarzanie, aby zmniejszyć czas obliczeń pod uwagę były brane tylko 2 płcie (MALE, FEMALE) oraz rasy (WHITE, BLACK). Po przeprowadzonej klasteryzacji program dokonywał dopasowania różnych kombinacji wartości etykiety utworzonej przez algorytm i porównywał to z prawdziwymi wartościami etykiety w rekordach. Dla przykładu, jeżeli dany eksperyment dotyczył płci, liczba klastrów była ustawiana na 2. Oznaczało to, że algorytm utworzy etykiety *0* oraz *1*. Następnie należało sprawdzić wszystkie permutacje etykiet względem unikatowych wartości. Przykład par wartość - etykieta z klastra: ('MALE',

0), ('FEMALE', 1) oraz ('MALE', 1), ('FEMALE', 0). Finalnie program wybiera układ w którym wartość *accuracy* jest największa. Podejście wiążę się ze znacznym wydłużeniem czasu działania programu w przypadku etykiet, które posiadają więcej niż 2 wartości (jak rasa). Wygenerowane wyniki zawierają informacje o wykorzystanym algorytmie, etykiecie, liczbie próbek poddanych analizie, permutacji, która osiągnęła najlepsze accuracy i wartość tej metryki.

6.3.2. Wyniki - Etykiety pojedyncze

Skrócone nazwy etykiet:
— M - male
— F - female
— B - black
— W - white

Wiersze	Po odrzuceniu	Algorytym	LABEL: M	LABEL: F	Accuracy
7375993	1960061	Kmeans	0	1	50.018
7375993	1960061	Aglomeracyjny	0	1	53.915
7375993	1960061	DBSCAN	-	-	-

Tabela 7. Wyniki grupowania dla etykiety SUSP_SEX

Wiersze	Po odrzuceniu	Algorytym	LABEL: B	LABEL: W	Accuracy
500000	150924	Kmeans	0	1	50.075
500000	150924	Aglomeracyjny	0	1	59.17
500000	150924	DBSCAN	-	-	-

Tabela 8. Wyniki grupowania dla etykiety VIC_RACE

6.3.3. Wyniki - Etykiety z wielu kolumn

Skrócone nazwy wartości w przypadku etykiety SUSP_SEX_VIC_SEX:
— MM - male on male
— MF - male on female
— FM - female on male
— FF - female on female

Wiersze	Po odrzuceniu	Algorytym	MM	MF	FM	FF	Accuracy
1000000	284593	Kmeans	2	0	1	3	25.07
1000000	284593	Aglomeracyjny	0	3	2	1	32.84
1000000	284593	DBSCAN	-	-	-	-	-

Tabela 9. Wyniki grupowania dla etykiety SUSP_SEX_VIC_SEX

Skrócone nazwy wartości w przypadku etykiety SUSP_RACE_VIC_RACE:
— BW - black on black
— BB - black on white
— WB - white on black
— WW - white on white

Wiersze	Po odrzuceniu	Algorytym	BW	BB	WB	WW	Accuracy
500000	68693	Kmeans	0	2	1	3	25.23
500000	68693	Aglomeracyjny	3	0	2	1	35.35
500000	68693	DBSCAN	-	-	-	-	-

Tabela 10. Wyniki grupowania dla etykiety SUSP_RACE_VIC_RACE

6.3.4. Analiza wyników

DBSCAN został uruchomiony z domyślnymi parametrami. Spowodowało to, że utworzył tylko jeden klaster i umieścił w nim wszystkie elementy oznaczając je etykiety -1, co jest równe punktowi typu outlier.

W przypadku algorytmu KMeans podczas testowania można było zauważyć, że algorytm osiąga accuracy zbliżone do 1/n%, gdzie n jest liczbą klastrów. Może to sugerować, że algorytm nie działa poprawnie i generuje wyniki losowo.

Algorytm aglomeracyjny osiągnął bardziej zróżnicowane wyniki, jednak należy przeprowadzić dodatkową weryfikację w celu ustalenia czy są one miarodajne.

Dostosowywanie parametrów algorytmów w celu poprawy jakości klasteryzacji oraz analiza wyników będą celami na następne checkpointy

Literatura

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i