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 3

Spis treści

1. Wprowadzenie . . . 2

2. Podział obowiązków w zespole . . . 2

3. Charakterystyka zbioru danych . . . 2

4. Cele projektu . . . 2

4.1. Klasyfikacja rodzaju lub poziomu przestępstwa . . . 2

4.2. Regresja czasu trwania przestępstwa . . . 3

4.3. Grupowanie przestępstw na podstawie podzbiorów cech . . . 3

5. Schemat operacji związanych z przetwarzaniem i analizą danych 4

5.1. Cel 1 - Klasyfikacja . . . 4

6. Wstępne przetwarzanie danych . . . 4

7. Przetwarzanie i analiza danych . . . 5

7.1. Cel 1 - Klasyfikacja . . . 5

7.2. Cel 2 - Klasteryzacja . . . 8

Literatura . . . 8

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 oraz rewizji na podstawie wyników uzyskanych w Checkpoint 2 zostały przedstawione w sekcji 4.

2. Podział obowiązków w zespole

— Szymon Gruda - Kontynuacja prac nad celem, związanym z klasyfikacją kodu przestępstwa.
— Jan Karwowski - Badanie sensowności dodanie OneHotEncoding oraz OrdinalEncoding, rozwiązywanie problemów z przepełnianiem się pamięci RAM po dokonaniu encodingu
— Michał Kidawa - kontynuacja badania eksperymentów związanych z grupowaniem elementów zbioru na podstawie istniejących oraz tworzonych cech
— Kamil Kowalewski - Zmiany w preprocessingu aby była możliwość osiągnięcia lepszy efektów niż te uzyskane w celach w Checkpoint 2 oraz przygotowania sprawozdania

3. Charakterystyka zbioru danych

Kolumny w zbiorze danych zdecydowaliśmy się 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 wy-

kroczenia, 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. Schemat operacji związanych z przetwarzaniem i analizą danych

5.1. Cel 1 - Klasyfikacja

Na podstawie niskiej dokładności klasyfikacji, wykorzystującej naiwny klasyfikator Bayes'a, zrezygnowano z dalszego testowania wykorzystania tego klasyfikatora, z powodu braku przesłanego, jakoby mógł spełniać swoje zadanie, bazując na danych z badanego zbioru danych. W celu poprawy rezultatów otrzymywanych na skutek działania klasyfikatora, poprawiono kodowanie wartości takich kolumn danych jak płeć, grupa wiekowa, dzień tygodnia i dzień roku. Następnie przeprowadzono serię eksperymentów, które pozwoliły określić czy cel w obecnej postaci jest możliwy do zrealizowania.

6. Wstępne przetwarzanie danych

W ramach wstępnego przetwarzania danych został przygotowany skrypt o nazwie *dataset_preprocessing.py*. Jego dokładne działanie zostało opisane w ramach Checkpoint 2 w sekcji o tej samej nazwie. Co więcej w ramach Checkpoint 3 zostały dokonane następujące zmiany aby dane były lepiej przygotowane do poszczególnych celów:

1. Konwersja Pandas Datetime na Timestamp aby były to liczby
2. Dodanie enkodowania kolumn z użyciem OneHotEncoder dla kolumn typu *categorical* - dzięki temu powstaję wiele nowych kolumn
3. Dodanie enkodowania kolumn z użyciem OrdinalEncoder dla kolumn typu *ordinal*
4. Zapis otrzymanego pliku po dokonaniu enkodowania wskazanymi wyżej metodami oraz dwukrotne wykorzystanie PCA na uzyskanym DataFrame z parametrem *n_components* równym liczbie kolumn oraz dwukrotności po pierwszej fazie preprocessingu tzn przed dokonaniem enkodowania. Wszystkie trzy pliki były dostępne do użycia w algorytmach i metodach Data Science do zrealizowania celów.

Po wykonaniu w/w skryptu okazała się niestety, że z powodu tego powstałej liczby kolumn nie jest możliwe przeprowadzanie analizy danych na całym pliku. Były dokonywane próby optymalizacji tego poprzez dokonanie encodingu w czasie działania programu bez tworzenia tak ogromnego pliku natomiast żaden z członków grupy nie dysponuję tak potężnym sprzętem, który miałby 40GiB RAMu. Problem ten zostanie omówiony z prowadzącym zajęcia w czasie przedstawiania tego Checkpointu celem otrzymania sugestii jak go rozwiązać. Dokonane różne warianty preprocessingu miały na celu polepszyć wyniki z Checkpoint 2 lecz niestety z przyczyn technicznych nie udało się tego dokonać.

7. Przetwarzanie i analiza danych

7.1. Cel 1 - Klasyfikacja

W ramach tego celu zmieniono kodowanie kolumn, przechowujących informację o rasie, płci, grupie wiekowej. Zarówno dla podejrzanego jak i dla ofiary przestępstwa. Następnie zakodowano status popełnionego przestępstwa. Z analizy danych przeprowadzonych w ramach Checkpoint'u 1, można było wnioskować, że dzień tygodnia oraz dzień roku (w kontekście pory roku) może przenosić ważne dla klasyfikacji informacje. Dlatego, aby zminimalizować straty informacji, związane z wykorzystaniem daty popełnienia przestępstwa. Dzień tygodnia oraz dzień roku, zakodowano jako parę wartości funkcji trygonometrycznych sinus i cosinus, dzięki czemu udaje się przechować informację o cyklicznej naturze czasu. Po przeprowadzeniu tych operacji przeprowadzono naukę oraz test klasyfikatora, wykorzystując do tego jeden milion próbek ze zbioru.

Nazwa parametru	Wartość parametru	Skuteczność trening	skuteczność test
min_samples_leaf	10	0.7065	0.6067
min_samples_leaf	100	0.6468	0.6342
min_samples_leaf	1000	0.6214	0.6183
min_samples_leaf	10000	0.5909	0.5909
max_depth	5	0.6006	0.6002
max_depth	10	0.6287	0.6248
max_depth	15	0.6673	0.6311
max_depth	25	0.8303	0.5875
n_estimators	100	0.6333	0.6287
n_estimators	1000	0.6343	0.6297
max_samples	0.01	0.5375	0.5370
max_samples	0.02	0.5697	0.5697
max_samples	0.05	0.6034	0.6030
max_samples	0.1	0.6162	0.6151
max_samples	0.2	0.6213	0.6194
max_samples	0.5	0.6295	0.6261
max_samples	0.8	0.6327	0.6284
max_samples	0.99	0.6344	0.6299

Tabela 1. Wyniki dla lasu losowego przy wykorzystaniu różnych parametrów, dla kodu przestępstwa

Następnym krokiem było wybranie najlepszych parametrów i zbadanie dla nich kilku metryk, dla dokładności klasyfikacji 62,86% wartości *Sensitivities* dla kodów przestępstw, prezentowały się tak: [0. 0.12 0.3734 0.5475 0.2113 0.7401 0. 0. 0. 0. 0. 0. 0.0389 0.9363 0.0721 0. 0.0574 0. 0. 0. 0.465 0. 0. 0. 0.0599 0.9807 0. 0. 0. 0. 0.822 0. 0. 0.9061 0. 0. 0. 0. 0. 0.1015 0. 0. 0. 0. 0. 0. 0.0098 0. 0.1207 0. 0. 0. 0. 0. 0.9989 0. 0. 0.]

Natomiast wartości *Precisions* [0. 0.4049 0.4818 0.4048 0.4834 0.4861 0. 0. 0. 0. 0. 0. 0.641 0.3771 0.4219 0. 0.4795 0. 0. 0. 0.3985 0. 0. 0. 0.6766 0.4683 0. 0. 0. 0. 0.7215 0. 0. 0.4829 0. 0. 0. 0. 0. 0.6505 0. 0. 0. 0. 0. 0. 0.9167 0. 0.472 0. 0. 0. 0. 0. 0.9954 0. 0. 0.]

Po zapoznaniu się z powyższymi wynikami oraz tabelą 1 można stwierdzić, że poczynione zmiany w celu poprawy dokładności klasyfikacji kodu przestępstwa nie zaskutkowały wielkimi zmianami. Jest to spowodowane najprawdopobniej niemożliwością wykonania takiej klasyfikacji bazując na tych danych. Dlatego zbadane zostało jak zachowuje się klasyfikator, przy klasyfikacji poziomu przestępstwa, wyniki eksperymentów zaprezentowano poniżej,

Nazwa parametru	Wartość parametru	Skuteczność trening	skuteczność test
min_samples_leaf	10	0.9999	0.9999
min_samples_leaf	100	0.9999	0.9999
min_samples_leaf	1000	0.9999	0.9999
min_samples_leaf	10000	0.9999	0.9999
max_depth	5	0.9999	0.9999
max_depth	10	0.9999	0.9999
max_depth	15	0.9999	0.9999
max_depth	25	0.9999	0.9999
n_estimators	100	0.6333	0.6287
n_estimators	1000	0.6343	0.6297
max_samples	0.01	0.9999	0.9999
max_samples	0.02	0.9999	0.9999
max_samples	0.05	0.9967	0.9968
max_samples	0.1	0.9991	0.9989
max_samples	0.2	0.9999	0.9999
max_samples	0.5	0.9999	0.9999
max_samples	0.8	0.9999	0.9999
max_samples	0.99	0.9999	0.9999

Tabela 2. Wyniki dla lasu losowego przy wykorzystaniu różnych parametrów, dla poziomu przestępstwa

Dla najlepszych parametrów zbadano metryki, dla dokładności klasyfikacji 99,99% wartości *Sensitivities* dla kodów przestępstw, prezentowały się tak: [0.9999 1. 0.9999]

Natomiast wartości *Precisions* [0.9999 0.9999 1.]

Macierz pomyłek (ang. *Confusion Matrix*) została zaprezentowana w tabeli 3, niestety dla kodu przestępstwa nie udało się takiej macierzy przedstawić w sposób czytelny, dlatego pominięto ją w sprawozdaniu.

VIOLATION	MISDEMEANOR	FELONY
35959	2	0
3	84691	1
0	3	46626

Tabela 3. Macierz pomyłek dla rodzaju przestępstwa

Wnioskując na podstawie rozbieżności pomiędzy dokładnościami klasyfikatorów, klasyfikujących kod przestępstwa a jego rodzaj można wywnioskować, że jeden cel jest zbyt ambitny, a drugi ma potencjał na uogólnienie. Dlatego w kolejnej iteracji badawczej, spróbujemy osiągnąć pewną modyfikację celu i wyznaczyć mniejszy zbiór kolumn (lub mniej oczywisty ich dobór), który pozwoli klasyfikować rodzaj popełnionego przestępstwa z zachowaniem możliwie największej dokładności klasyfikacji.

7.2. Cel 2 - Klasteryzacja

Pomimo zmian zarówno w parametrach klasteryzacji, jak i charakterystyce zbioru, który po preprocessingu trafił do metod związanych z grupowanie nie udało się poprawić jakości klasteryzacji. Przy zmianie liczby klastrów otrzymywane wyniki grupowania nadal stanowiły trend podziału danych na 1 / N, gdzie N to liczba klastrów. Algorytm DBSCAN nadal klasyfikował wszystkie próbki jako elementy wyjątkowe, czyli outliers.

Literatura

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