Data 03.06.2024

Małgorzata Mokwa
Sebastian Pergała
Zespół nr 4

Raport walidacyjny

Projekt 2 - klasteryzacja
Walidacja zespołu 6

Spis treści

1 Wstęp

Raport walidacyjny został sporządzony przez zespół numer 4, tj. Małgorzatę Mokwę i Sebastiana Pergałę. Walidowanym zespołem był zespół numer 6, składający się z Julii Kruk i Michała Piechoty.

Celem projektu wykonywanego przez zespół walidowany było skonstruowanie modelu uczenia nienadzorowanego i klasteryzacja określonych danych.

Zbiór danych, na którym zespół walidowany pracował, pochodzi z witryny internetowej *Kaggle* : *Galaxy Zoo 2: Images* (`https://www.kaggle.com/datasets/jaimetrickz/galaxy-zoo-2-images`).

Celem biznesowym, postawionym przez zespół walidowany było grupowanie obrazów galaktyk ze zbioru danych na podstawie ich podobieństwa za pomocą stworzonego modelu. W tym celu zespół przedsięwziął się na wykonanie pracy, w której można wyróżnić trzy etapy:

1. Eksploracyjna analiza danych.
2. Inżynieria cech.
3. Modelowanie i walidacja modeli.

2 Eksploracyjna analiza danych

Zbiór danych składał się z 243434 obrazów typu *.jpg* o rozmiarach 424×424. Obrazy były nazwane kolejnymi liczbami całkowitymi (1.jpg, 2.jpg, ...).

Jednym z pomysłów zespołu walidowanego w celu pozyskania lepszych danych do określonego wcześniej celu było skorzystanie z faktu umiejscowienia galaktyk na środkach zdjęć i ucięcie brzegów obrazów, co miało zmniejszyć wymiarowość danych i pozbycie się niepotrzebnych informacji. W rzeczywistości takie podejście jednak nie brało pod uwagi istnienia obrazów, gdzie galaktyka pokrywała większy obszar i opisana obróbka skutkowałaby zniekształceniem danych.

Wystarczył szybki wgląd w katalog z obrazami galaktyk, aby czujne oko walidatorów znalazło kilka przykładów, gdzie galaktyki były znacznie większe, niż te wylosowane w małej próbce danych przez zespół walidowany podczas analizy danych, a kontynuowanie używania metody ucinania brzegów zdjęć wiązałoby się z usunięciem znaczącej części galaktyki z obrazu.

Brakowało również analizy kolorów i jasności zdjęć, gdzie również była możliwość wyciągnięcia wartościowych informacji na temat struktury danych.

Zgłoszone przez zespół walidujący zostały następujące uwagi:

1. Brak pliku requirements.txt z wersjami użytych bibliotek.
2. Na podstawie 10000 pierwszych zdjęć stweirdzono, że wszystkie mają ten sam rozmiar, co nie musi być prawdą.
3. Ucinanie brzegów zdjęć nie wydaje się być dobrym pomysłem. Niechciane obiekty w tle, niebędące galaktykami, dalej mogą znajdować się na pozostawionym obrazie i zwiększone jest ryzyko, że ucięta zostanie część galaktyki (jak to jest w przypadku zdjęć np. 185281.jpg, 188580, 191118, 14242, 15099, 191447, ...).
4. Niektóre zdjęcia różnią się znacząco od pozostałych (np. 181121, 8720, 13792, 183816, 194203, ...) dobrze by było zwrócić uwagę na potencjalne outliery.
5. Warto by było zbadać rozkład zdjęć ze względu na jasność. Niektóre zdjęcia są znacząco jaśniejsze od pozostałych, a ich tło nie jest czarne (np. 215958 w porównaniu do 217917). Proponumeny też zrobienie analizy kolorów galaktyk, identyfikacjy najczęściej występujących barw galaktyk i analizy wielkości lub ilości galaktyk, na paru wygenerowanych obrazach jest więcej niż jeden obiekt. Dobrze to widać po dodaniu thresholdu. Można by policzyć ilość obiektów i zobaczyć jak się to rozkłada w całym zbiorze.

Komentarze do dotychczas napisanego kodu były przyjęte pozytywnie przez zespół walidowany, a na prezentacji kamienia milowego z etapu eksploracjnej analizy danych można było zauważyć wdrożone do projektu poprawki, świadczące o poważnym podejściu do wykonywanego zadania i braku lekceważenia nieomylnego zdania walidatorów.

3 Inżynieria cech

W fazie inżynierii cech (Feature Engineering) zespół budujący zdecydował się na grupowanie galaktyk znajdujących się w centralnej części obrazów. Jedną z uwag zespołu walidującego na etapie Eksploracyjnej Analizy Danych była obserwacja, że na niektórych zdjęciach znajduje się więcej niż jedna galaktyka. Nie wiadomo było, czy zespół budujący będzie skupiał się tylko na środkowej czy wykorzysta wszystkie dostępne informacje. Sprecyzowanie tego na etapie FE rozwiało wątpliwości.

Kolejnym plusem było zastosowanie różnych metod ekstrakcji cech, w tym różnych modeli CNN oraz histogramu gradientów (HOG). Wykorzystanie wielu podejść, zamiast polegania na jednej metodzie, umożliwiło lepsze uchwycenie różnorodnych charakterystyk galaktyk.

Cały proces inżynierii cech został zrealizowany w sposób profesjonalny. Notatnik i kod zostały utrzymane w bardzo biznesowym stylu. Zawarto wiele komentarzy oraz wniosków w każdej części projektu. Bardzo to ułatwiło zrozumienie toku myślowego zespołu budującego oraz ułatwiło porównywanie wniosków i obserwacji naszych i zespołu budującego.

Podsumowując, faza inżynierii cech była kluczowym etapem projektu, który dzięki zastosowaniu odpowiednich technik i dbałości o szczegóły w dokumentacji, pozwolił na uzyskanie wysokiej jakości danych i przyczynił się do sukcesu całego procesu klasteryzacji obrazów.

4 Modelowanie i walidacja modeli

W fazie modelowania dużym plusem jest szczegółowa analiza decyzji dotyczącej optymalnej liczby klastrów, na które model powinien dzielić zdjęcia galaktyk. Zastosowano różne metody oceny liczby klastrów, w tym metodę łokcia (elbow method) oraz silhouette score. Pozwoliło na dokładne określenie odpowiedniej liczby klastrów przed przystąpieniem do tworzenia modelu.

Model KMeans został opracowany i omówiony bardzo szczegółowo. Zespół przeprowadził kompleksową analizę tego modelu, co jest dużym atutem projektu. Natomiast model DBSCAN został omówiony tylko pobieżnie. W przypadku DBSCAN zabrakło miar oceny modelu oraz wizualizacji zdjęć zakwalifikowanych do różnych klastrów.

W raporcie brakowało również podsumowania oraz wniosków, które wskazywałyby, który model jest najlepszy i dlaczego. Dodatkowo, zabrakło informacji na temat powstałych klastrów. Modele dzieliły zdjęcia na 3 grupy. Patrząc na zdjęcia i ich kategorie trudno jest zrozumieć, dlaczego jest taki podział. Wydaje nam się, że żadna z grup nie ma cech, które wyróżniałyby je na tle innych grup.

Warto byłoby również rozważyć ewaluację modeli za pomocą innych metryk niż tylko silhouette score. Przykładowo, mógłby to być wskaźnik Calinskiego-Harabasza.

Uwagi do części Feature Engineering oraz Modelowanie i Walidacja:

1. Dalej brakuje pliku requirements.txt z użytymi wersjami bibliotek.
2. W etapie eksploracji zostały znalezione obrazy, które mają np. większą średnią jasność od innych. Czy na tych zdjęciach metoda bounding boxes też działa?
3. Można było przetestować większą ilość metod pozyskiwania cech z obrazów.
4. Podobnie dobrze by było sprawdzić inne algorytmy oprócz KMeans i porównać ze sobą wyniki.
5. Jakość stworzonych klastrów należałoby zmierzyć za pomocą różnych metryk.
6. Na plus jest stworzenie biblioteki z funkcjami wykorzystywanymi w różnych notatnikach