Plan eksperymentu - uczenie maszynowe

Rafał Smolak (252973)
Hubert Gabory (255743)

Maj 2023

1 Research questions

1. Czy gęstościowe ważenie wzorców prowadzi do lepszych wyników uczenia sieci neuronowych dla zbiorów niezbalansowanych?
2. W jaki sposób stopień niezbalansowania zbiorów, wpływa na wyniki uczenia sieci neuronowych, przy zastosowaniu gęstościowego ważenia wzorców?
3. Czy gęstościowe ważenie wzorców daje lepsze rezultaty od pozostałych metod podejmujących problem niezbalansowania zbiorów danych?

2 Dokładne cele eksperymentów

1. Ocena wpływu gęstościowego ważenia wzorców opartego na jądrowej estymacji gęstości(KDE) na wyniki uczenia sieci neuronowych dla różnych niezbalansowanych zbiorów danych dla problemu klasyfikacji binarnej.
    - Eksperyment obejmujący uczenie sieci neuronowych na różnych niezbalansowanych zbiorach danych z i bez zastosowania gęstościowego ważenia wzorców.
    - Porównanie wyników na podstawie miar odpowiednich dla problemu niezbalansowanych danych, takich jak precyzja, czułość, czy miara F1.
2. Analiza wpływu stopnia niezbalansowania zbiorów na wyniki uczenia sieci neuronowych z zastosowaniem gęstościowego ważenia wzorców dla różnych zbiorów danych w klasyfikacji binarnej.
    - Przeprowadzenie serii eksperymentów z różnymi stopniami niezbalansowania zbiorów danych (np. 2:1, 5:1, 10:1) z zastosowaniem gęstościowego ważenia wzorców.
    - Porównanie wyników uczenia sieci neuronowych na różnych poziomach niezbalansowania, aby zidentyfikować potencjalne zależności między stopniem niezbalansowania a efektywnością gęstościowego ważenia wzorców.
3. Porównanie gęstościowego ważenia wzorców z innymi metodami radzenia sobie z problemem niezbalansowania zbiorów danych w kontekście uczenia sieci neuronowych dla różnych zbiorów danych.
    - Wybranie kilku innych technik radzenia sobie z niezbalansowanymi danymi, takich jak random oversampling, SMOTE, czy ADASYN.
    - Przeprowadzenie eksperymentów, w których uczenie sieci neuronowych zostanie przeprowadzone przy użyciu różnych metod radzenia sobie z niezbalansowanymi danymi.
    - Porównanie wyników eksperymentów dla poszczególnych metod w celu oceny, czy gęstościowe ważenie wzorców daje lepsze rezultaty. Porównanie powinno być oparte na miarach takich jak precyzja, czułość, miara F1.

3 Zbiory danych

Wybrane zbiory danych to zbiory klasyfikacji binarnej. Oznacza to, że wynikiem klasyfikacji jest liczba binarna 0 albo 1, która oznacza wystąpienie lub brak wystąpienia danego zjawiska. Wybrane zbiory danych są zbiorami zbalansowanymi, ale będą ręcznie przekształcane w zbiory niezbalansowane, o różnych stopniach niezbalansowania w celu kompleksowej analizy wpływu stopnia niezbalansowania na efektywność testowanej przez nas metody ważenia wzorców. Poniżej przedstawiamy wybrane przez nas zbiory danych:

3.1 Cancer data

Zbiór danych wskazujących czy nowotwór jest łagodny czy złośliwy bazując na jego cechach.

- URL: https://www.kaggle.com/datasets/erdemtaha/cancer-data
- Cechy(30): id, **diagnosis**, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst
- Liczba instancji - 570

3.2 Diabetes prediction dataset

Zbiór danych wskazujących czy pacjent jest cukrzykiem bazując na jego cechach.

- URL: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
- Cechy(9): gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, **diabetes**
- Liczba instancji - 100000

3.3 League of Legends Diamond Games (First 15 Minutes)

Zbiór danych wskazujący, która drużyna wygra grę na podstawie statystyk z pierwszych piętnastu minut rozgrywki.

- URL: https://www.kaggle.com/datasets/benfattori/league-of-legends-diamond-games-first-15-minutes
- Cechy(18): matchId, **blue_win**, blueGold, blueMinionsKilled, blueJungleMinionsKilled, blueAvgLevel, redGold, redMinionsKilled, redJungleMinionsKilled, redAvgLevel, blueChampKills, blueHeraldKills, blueDragonKills, blueTowersDestroyed, redChampKills, redHeraldKills, redDragonKills, redTowersDestroyed
- Liczba instancji - 48651

4 Plan eksperymentu

1. Wykonanie preprocesingu zbiorów danych jeżeli będzie taka potrzeba (na przykład w przypadku brakujących lub wadliwych danych).
2. Przygotowanie środowiska eksperymentalnego w języku python, przy użyciu scikit-learn'a, matplotlib'a i innych bibliotek.

3. Dla każdego wybranego zbioru danych, przeprowadzenie eksperymentów z zastosowaniem gęstościowego ważenia wzorców opartego na jądrowej estymacji gęstości (KDE). Uczenie sieci neuronowych zostanie przeprowadzone na różnych niezbalansowanych zbiorach danych, z różnymi stopniami niezbalansowania (np. 2:1, 5:1, 10:1).

4. Dla każdego zbioru danych, przeprowadzenie eksperymentów z zastosowaniem innych technik radzenia sobie z niezbalansowanymi danymi, takich jak random oversampling, SMOTE czy ADASYN. Uczenie sieci neuronowych zostanie przeprowadzone na różnych niezbalansowanych zbiorach danych, z różnymi stopniami niezbalansowania (np. 2:1, 5:1, 10:1).

5. Dla każdego zbioru danych, przeprowadzenie eksperymentów bez zastosowania żadnych technik radzenia sobie z niezbalansowanymi danymi (tj. uczenie sieci neuronowych na surowych danych). Uczenie sieci neuronowych zostanie przeprowadzone na różnych niezbalansowanych zbiorach danych, z różnymi stopniami niezbalansowania (np. 2:1, 5:1, 10:1).

6. Wykorzystanie walidacji krzyżowej z podziałem na 5 zestawów (5-fold cross-validation) w celu obiektywnej oceny jakości uczenia się sieci neuronowych.

7. Wykorzystanie odpowiednich miar oceny jakości uczenia się sieci neuronowych, takich jak precyzja, czułość, czy miara F1, do porównania wyników eksperymentów z różnymi metodami radzenia sobie z problemem niezbalansowania zbiorów danych.

8. Sporządzenie tabel i wykresów prezentujących otrzymane wyniki badań i analizy statystycznej oraz wyciągnięcie wniosków.

5 Opis środowiska eksperymentalnego

Środowisko eksperymentalne zostanie zbudowane na podstawie języka Python, który umożliwia szeroką gamę narzędzi i bibliotek do pracy z uczeniem maszynowym i analizą danych. Dla realizacji projektu wykorzystane zostaną następujące narzędzia i biblioteki:

5.1 Scikit-learn

Scikit-learn został stworzony w celu zapewnienia prostego i jednolitego interfejsu dla różnych algorytmów uczenia maszynowego. Biblioteka ta dostarcza również narzędzi do przetwarzania danych, takich jak normalizacja, skalowanie i kodowanie cech.

W ramach projektu biblioteka scikit-learn zostanie wykorzystana do implementacji algorytmu oraz do oceny jakości tego algorytmu na zestawie danych. Dzięki temu będziemy mogli porównać różne modele.

5.2 NumPy

Jest to popularne narzędzie w analizie danych, szczególnie w przypadku dużych zbiorów danych, gdzie wydajność i szybkość obliczeń są kluczowe.

W ramach projektu biblioteka NumPy zostanie wykorzystana do przetwarzania danych oraz obliczeń numerycznych. Z uwagi na to, że pracujemy z dużymi zbiorami danych, NumPy będzie przydatny do przetwarzania i organizowania danych w postaci tablic wielowymiarowych. Dzięki temu, będziemy w stanie łatwo wykonywać operacje matematyczne na danych, takie jak obliczanie statystyk opisowych, czy też normalizacja danych.

5.3 Pandas

Biblioteka pandas zapewnia wiele funkcji do manipulacji i filtrowania danych, w tym do grupowania, łączenia, wypełniania braków i filtrowania danych. Pandas pozwala również na łatwe tworzenie wykresów i wizualizacji danych.

W ramach projektu biblioteka pandas zostanie użyta do zarządzania danymi oraz ich wstępnego przetwarzania. Pandas pozwoli na łatwe wczytywanie i manipulowanie danymi, takie jak usuwanie duplikatów, uzupełnianie brakujących danych oraz przeprowadzanie operacji matematycznych. Dzięki temu, będziemy w stanie lepiej zrozumieć dane, przygotować je do analizy oraz przetestować algorytmy uczenia maszynowego.

5.4 Matplotlib

Matplotlib to biblioteka dla języka Python, która umożliwia tworzenie wykresów i wizualizacji danych. Matplotlib pozwala na tworzenie różnych typów wykresów, w tym wykresów liniowych, słupkowych, kołowych, histogramów, wykresów punktowych, a także wykresów konturowych i trójwymiarowych.

Biblioteka matplotlib jest bardzo elastyczna i oferuje wiele opcji konfiguracyjnych, które pozwalają na dostosowanie wyglądu wykresów do indywidualnych potrzeb. Matplotlib pozwala na dodawanie tytułów, osi, etykiet, legend i innych elementów, co ułatwia interpretację danych i prezentację wyników.

W ramach projektu biblioteka matplotlib zostanie wykorzystana do prezentacji wyników badań oraz analizy statystycznej. Matplotlib pozwoli na wizualizację danych, co ułatwi analizę wyników eksperymentów oraz przedstawienie wniosków w przejrzysty i zrozumiały sposób. Matplotlib będzie również wykorzystywany do tworzenia wykresów porównawczych, które pozwolą na porównanie wyników uzyskanych dla różnych algorytmów uczenia maszynowego.

5.5 Scipy.stats

Biblioteka scipy.stats to moduł statystyczny z biblioteki SciPy, który umożliwia przeprowadzanie różnego rodzaju testów statystycznych oraz analizę danych. Biblioteka ta zawiera funkcje umożliwiające obliczanie wartości p-wartości, estymowanie parametrów rozkładów, wykonanie testów hipotez i wiele innych.

W ramach projektu biblioteka scipy.stats zostanie wykorzystana do przeprowadzenia testów parowych oraz globalnych rankingowych. Testy parowe pozwalają na porównanie wyników uzyskanych dla dwóch różnych algorytmów uczenia maszynowego na tym samym zestawie danych. Testy globalne rankingowe pozwalają na porównanie wyników uzyskanych dla kilku algorytmów uczenia maszynowego i wyłonienie najlepszego z nich.

Biblioteka scipy.stats umożliwia również przeprowadzenie analizy regresji, testów normalności, testów jednorodności, testów niezależności oraz wiele innych testów statystycznych, które mogą być przydatne w analizie danych i wyników eksperymentów związanych z uczeniem maszynowym.