10.12.2024

Projekt - przedmiot UMA

Zespół w składzie:
Jakub Bąba
Adrian Murawski

Treść zadania

Przygotować implementację algorytmu do indukcji reguł [IRep++] oraz przeprowadzić eksperymenty na wybranym zadaniu z bazy.

Opis problemu i jego sposobu rozwiązania

Zadanie ma na celu implementację algorytmu IRep++ oraz przeprowadzenie eksperymentów, które pokażą działanie i wyniki tego algorytmu. Opis algorytmu, wraz z jego działaniem zamieściliśmy poniżej.

Planujemy zaimplementować algorytm w Pythonie, a eksperymenty przeprowadzić i udokumentować przy pomocy *notebooków pythona*. Takie *notebooki* pozwolą zachować rezultat uruchomienia kodu, czyli w podanym przypadku wyniki eksperymentów.

Do eksperymentów planujemy użyć zbioru dotyczącego klasyfikacji raka piersi - jako złośliwego, bądź też łagodnego. Eksperymenty, które przeprowadzimy na tym zbiorze są opisane w sekcji “Szacunkowy plan eksperymentów”.

Opis algorytmu IRep++

Algorytm IREP++ jest algorytmem indukcji reguł decyzyjnych. Buduje on reguły decyzyjne oraz przycina je bazując na estymacji błędu rzeczywistego (na podstawie zbioru walidacyjnego). Algorytm ten bazuje swoje działanie na algorytmach RIPPER oraz IREP (Induced Reduced Error Pruning), jednak usprawnia proces tworzenia reguł, dbając o ich odpowiednią ogólność oraz minimalizując ryzyko nadmiernego dopasowania.

Zaprojektowany jest on do klasyfikacji binarnej, czyli do używania na danych, których funkcja celu przyjmuje dwie wartości.

Algorytm działa następująco:

1. Dzieli zbiór treningowy na dwie części:
    - grow set - wykorzystywany do tworzenia reguł,
    - prune set - wykorzystywany do przycinania reguł.
2. Tworzy pustą regułę, ustawiającą wartość funkcji celu na jedną z wartości.
3. Wykonuje kroki aż wszystkie przypadki z grow set zostaną sklasyfikowane lub nie będzie można tworzyć nowych reguł:
    a. Buduje reguły - uszczegóławia i rozszerza istniejącą regułę o dodatkowe warunki bazując na zbiorze grow set.

b. Przycina reguły - sprawdza nową regułę na zbiorze prune set oraz usuwa te warunki, które nie poprawiają wydajności, pozbywając się nadmiernego dopasowania.
c. Usuwa przykłady spełniające warunki z grow set (w celu skupienia się na pozostałych przypadkach).

Szacunkowy plan eksperymentów

Planujemy wykonać następujące eksperymenty w celu dokładnego zbadania, jak działa powyższy algorytm:

- Porównanie algorytmów IREP++, RIPPLE oraz IREP na zbiorze danych
- Modyfikacja rozmiaru zbioru treningowego i wpływ takiej modyfikacji na dokładność
- Modyfikacja stosunku rozmiaru zbioru budującego drzewo, do tego je ograniczającego (domyślnie jest to ⅔ do ⅓) i sprawdzenie dokładności
- Odporność algorytmu na szum - zmodyfikowanie części (np. 20% danych) poprzez losową zmianę wartości cech lub klas, a następnie sprawdzenie, jak wpływa to na dokładność klasyfikacji
- Przydatność poszczególnych kolumn do ostatecznego wyniku (przykładowo: usunięcie jednej kolumny - cechy - i sprawdzenie jak wpłynie to na dokładność klasyfikacji)

Zbiór danych do badań

W celu rozwiązania naszego problemu planujemy użyć zbioru danych dotyczących raka piersi (dane ze strony kaggle.com). Naszym celem jest klasyfikacja typu nowotworu jako złośliwy (M - *malignant*) lub łagodny (B - *benign*). Funkcja celu jest binarna, co odpowiada wymaganiom naszego algorytmu.

Zbiór danych zawiera 569 rekordów, a każdy rekord zawiera ID rekordu, funkcję celu oraz 30 parametrów numerycznych opisujących dane medyczne, takie jak promień czy wklęsłość.

Przygotowanie zbioru danych będzie polegało na pozbyciu się kolumny z ID, standaryzacji parametrów liczbowych oraz odpowiednim podziale danych na zbiór treningowy i testowy.

Źródła

https://www.researchgate.net/publication/220907085_IREP_a_Faster_Rule_Learning_Algorithm
http://elektron.elka.pw.edu.pl/~pcichosz/uma/slajdy/uma-s9.pdf
https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset