Politechnika Wrocławska
Wydział Elektroniki
Uczenie maszyn

Porównanie metod selekcji cech dla wybranych typów klasyfikatorów

Termin zajęć:
Środa TP 15:15
Autorzy:
Jakub Chmiel 235028
Janusz Długosz 235746
Prowadzący:
mgr inż. Paweł Zyblewski

7 grudnia 2020

1 Założenia projektowe

Celem projektu jest porównanie metod selekcji dla wybranych typów klasyfikatorów pod kątem trafności klasyfikacji. Projekt został wykonany w języku Python z wykorzystaniem biblioteki *scikit-learn*[1]. Na podstawie otrzymanych wyników zostanie przeprowadzona ich analiza w formie testu statystycznego. Na potrzeby wykonania testów statystycznych wykorzystano bibliotekę *scipy*[2].
Wybrano następujące metody selekcji cech:

- SelectKBest - polega na sporządzeniu rankingu cech, a następnie wybraniu określonej liczby najlepszych cech,
- VarianceThreshold - polega na wykluczeniu cech, których wartości nie różnią się znacząco w ramach zbioru danych,
- RFE (Recursive Feature Elimination) - wykorzystując funkcję estymatora rekurencyjnie eliminowane są cechy uznawane za mniej znaczące.

Wybrano następujące klasyfikatory:

- KNeighborsClassifier - polega na znalezieniu k próbek o najbardziej podobnym zestawie cech. Próbce przypisywana jest klasa najczęściej pojawiająca się w zbiorze sąsiadów,
- PassiveAggressiveClassifier - polega na sekwencyjnej analizie danych. Model analizuje kolejne próbki jedna za drugą. Jeśli jego prognoza jest poprawna, model pozostaje niezmieniony, w przeciwnym razie model jest dostosowywany,
- MLPClassifier - najpopularniejszy typ sieci neuronowych. Sieć składa się z jednej warsty wejściowej, jednej wyjściowej i wielu warstw ukrytych.
- ComplementNB - klasyfiaktor z grupy Naive Bayes, bazujący na twierdzeniu Bayesa. Polega na policzeniu prawdopodobieństwa przynależności próbki do każdej z klas,

2 Plan eksperymentu

Do celów eksperymentu wybrano 3 metody selekcji cech i 4 klasyfikatory. W celu wydzielenia zbiorów trenujących i testowych wykorzystano 2-krotną walidację krzyżową ze stratyfikacją i 5-krotnym powtórzeniu. W każdym porównaniu wybierana będzie metoda selekcji cech oraz 2 spośród wybranych klasyfikatorów, zatem łącznie wykonane zostanie 18 porównań. Dla każdego z klasyfikatorów w ramach porównania wykonane zostanie 10 prób, które zostaną wykorzystane jako ocena modelu.

Ostatnim etapem będzie wykonanie testów statystycznych. Dla porównania klasyfikatorów został wykorzystany test T-studenta[3]. Dla testu przyjęto wartość parametru $\alpha = 0,05$.

Eksperyment zostanie wykonany na zbiorze danych pochodzącym z repozytorium UCI[4], w którym przedstawione są oceny win białych w skali 1-10. Każda z 4898 instancji opisana jest przez 11 cech odpowiadających właściwościom fizycznym i chemicznym.

Bibliografia

[1] https://scikit-learn.org/stable/index.html

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

[3] https://pogotowiestatystyczne.pl/slowniczek/test-t-studenta/

[4] https://archive.ics.uci.edu/ml/datasets/Wine+Quality