BAGGING

GRZEGORZ BORKOWSKI

Wstęp

Zbiorem danych jakim się zajmowałem jest zbiór o nazwie "Adult" zwany również jako "Census Income" [1] dostępny w UCI Machine Learning Repository. Zadaniem była klasyfikacja binarna, należało przewidzieć mając dane wskaźniki ekonomiczno-społeczne danej osoby, czy ta osoba zarabia powyżej 50 tysięcy dolarów rocznie.

Zbiór treningowy składał się z 32561 próbek, zbiór testowy z 16281 próbek. Autor w pracy [2] osiągneli skuteczność klasyfikatora 84.47% przy pomocy hybrydowego klasyfikatora NBTree (hybryda klasyfikatora Naive Bayes i Decision Tree).

Rozwiązanie

Po wstępnym czyszczeniu danych, podzieliłem zbiór treningowy na 5 równolicznych zbiorów, każdy składał się z 80% danych treningowych (analogicznie jak w metodzie KFold).

Trenowałem pięć różnych klasyfikatorów, każdy klasyfikator trenowałem na jednym z pozdbiorów zbioru treningowego, a jako zbiór cross-validation, brałem pozostałe 20% danych treningowych, aby odpowienio dobrać parametry klasyfikatora.

Wybrane przeze mnie klasyfikatory to: dwa klasyfikatory oparte o metodę K najbliższych sąsiadów, klasyfikator oparty o drzewa decyzyjne, liniowy SVM i SVM z kernel rbf.

	KNN 1	Decision Tree	Linear SVM	SVM Kernel	KNN 2
Skuteczność	0.7973	0.8536	0.8252	0.7637	0.7958
Parametry	k=10, minkowski	max_depth=8, max_features=11	C=1.0, metric='l1'	kernel='rbf', C=0.001	k=8

Tabela 1: Opis użytych klasyfikatorów wraz z ich skutecznością na zbiorze testowym i parametrami

Stworzyłem prosty klasyfikator głosujący, przypisujący każdemu klasyfikatorowi równą wagę, a rezultatem była klasa, która uzyskała najwięcej głosów. Wybrałem pięć klasyfikatorów, aby uniknąć remisów w głosowaniu. Skuteczność jaką uzyskałem to: 0.8114.

Nastepnie powtórzyłem wyniki, biorąc klasyfikator KNN 1 z wagą głosu równą 1, Decision Tree z wagą równą 2, Linear SVm z wagą równą 1, KNN 2 z wagą równą 1 i osiągnałem skuteczność 0.840.

Wnioski

(1) Klasyfikator Decision Tree w tym problemie osiągą z łatwością bardzo wysoką skuteczność.
(2) Dla większej liczby próbek nie jest możliwe dokłane przetestowanie i wybranie optymalnych parametrów dla klasyfikatora SVM z kernelem nieliniowym na komputerze osobistym. Czas treningu wahał się od kilkunastu minut do kilku godzin, co uniemożliwiło mi odpowiedni dobór parametrów, co może tłumaczyć stosunkową niską skuteczność SVM z kernelem nieliniowym.
(3) Bagging to bardzo interesująca metoda, pomimo, że w tym przypadku nie osiągnęła skuteczności wyższej niz najlepszy z klasyfikatorów, to warto ją testować i korzystać, aby ograniczyć błędy poszczególnych klasyfikatorów i osiągnąć większa pewność w skuteczność stworzonego modelu.

Bibliografia

[1] https://archive.ics.uci.edu/ml/datasets/adult
[2] Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid"