Auto ML HW 1

Paweł Gelar

1 Wstęp

Celem zadania było zbadanie tunability wybranych algorytmów uczenia maszynowego korzystając z różnych zbiorów danych.

2 Opis eksperymentu

2.1 Zbiory danych

Eksperymenty przeprowadzono na czterech zbiorach danych dotyczących klasyfikacji binarnej z portalu OpenML[1]:

1. datatrieve (ID 1075),
2. tic-tac-toe (ID 50),
3. diabetes (ID 37),
4. pc1 (ID 1068)

Nie zawierały one braków danych. Szczegóły znajdują się w tabeli 1. Przed trenowaniem modeli zmienne kategoryczne zostały przetransformowane za pomocą one-hot encodera, a zmienne ciągłe za pomocą standard scalera. Zbiory zostały podzielone na część treningową i testową w proporcji 8:2.

2.2 Wyznaczenie optymalnych wartości hiperparametrów

Użyto przeszukiwania losowego i bayesowskiego. W obu przypadkach przeprowadzono 50 iteracji na 5-krotnej kroswalidacji. Rozkłady hiperparametrów przedstawiono w tabeli 2. Starano się je dobrać by były jak najbardziej zbliżone do tych przedstawionych w artykule [1]. Dzięki zastosowaniu stałego ziarna w przeszukiwaniu losowym testowane wartości hiperparametrów były stałe dla każdego zbioru. Obliczono globalnie optymalne hiperparametry jako te, które miały największą średnią dokładność.

3 Wyniki

3.1 Porównanie metod przeszukiwania

Wykresy przedstawiające porównanie przebiegów przeszukiwania przedstawiono na rysunku 1. Można zauważyć, że optymalizacja bayesowska nigdy nie osiągnęła ostatecznie gorszego wyniku, niż przeszukiwanie losowe. Jedynie mogło jej to zająć trochę dłużej. Ogółem wyniki tych przeszukiwań są podobne. Różnica mogłaby się powiększyć wraz z liczbą iteracji.

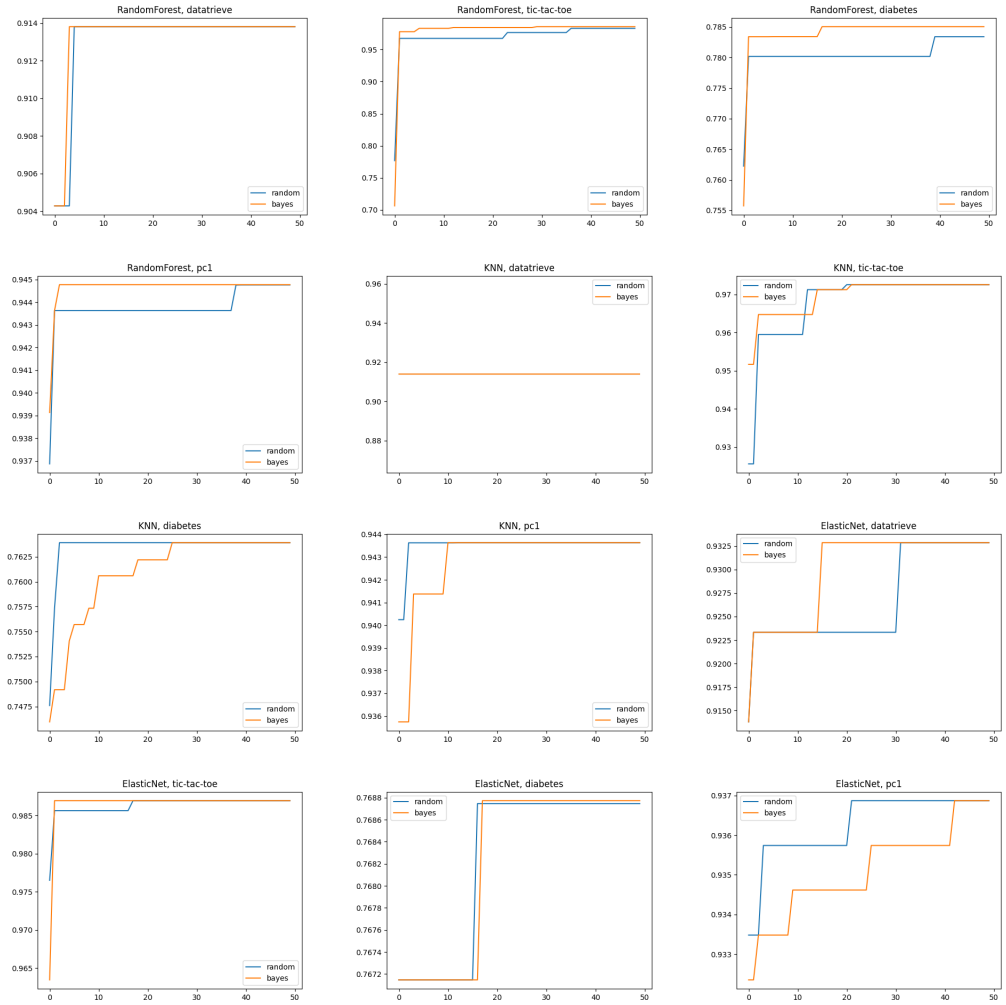
[1] www.openml.org

Nazwa	Liczba rekordów	Liczba kolumn kategorycznych	Liczba kolumn ciągłych
datatrieve	130	9	0
tic-tac-toe	958	0	10
diabetes	768	10	0
pc1	1109	22	0

Tabela 1: Szczegóły wykorzystanych zbiorów danych

Algorytm	Hiperparametr	Typ	Maks	Min	rozkład
Las losowy	n estimators	liczba całkowita	1	2000	jednostajny
	max samples	ułamek	0.1	1	jednostajny
	max features	ułamek	0	1	jednostajny
	min samples split	ułamek	0.001	0.2	logarytmiczno-jednostajny
KNN	n neighbors	liczba całkowita	1	31	jednostajny
	weights	kategoria	-	-	(uniform, distance)
ElasticNet	C	ułamek	0.0001	1	jednostajny
	l1 ratio	ułamek	0	1	jednostajny

Tabela 2: Rozkłady z których próbkowano wartości hiperparametrów. Wartości typu 0.0001 wynikają z braku możliwości zdefiniowania przedziałów otwartych przy optymalizacji Bayesowskiej.



Rysunek 1: Przebieg przeszukiwań (skumulowany najlepszy wynik). Kolorem niebieskim oznaczono przeszukiwanie losowe, a pomarańczowym bayesowskie

3.2 Tunowalność modeli

Tabela 3 przedstawia różnicę w wynikach testowych modeli z różnymi parametrami. W ponad połowie przypadków wynik najlepszy dla danego zbioru był równy temu dla globalnie optymalnych hiperparametrów. Najbardziej znaczącą różnicą było 3% poprawy dla lasu losowego. Większą różnicę dało się zauważyć między parametrami domyślnymi dla pakietu scikit-learn, a resztą.

Model	**Dataset**	**Różnica wyników (Best - Specific)**	**Różnica wyników (Best - Default)**
RandomForest	datatrieve	0.000%	0.000%
RandomForest	tic-tac-toe	0.000%	1.042%
RandomForest	diabetes	-3.247%	0.649%
RandomForest	pc1	0.450%	0.450%
KNN	datatrieve	0.000%	3.846%
KNN	tic-tac-toe	0.000%	5.729%
KNN	diabetes	-0.649%	-0.649%
KNN	pc1	0.000%	0.000%
ElasticNet	datatrieve	0.000%	0.000%
ElasticNet	tic-tac-toe	-0.521%	-0.521%
ElasticNet	diabetes	0.000%	0.000%
ElasticNet	pc1	0.450%	0.000%

Tabela 3: Porównanie wyników testowych modeli z różnymi hiperparametrami (w punktach procentowych). 'Best' oznacza hiperparametry, które były średnio najlepsze dla wszystkich zbiorów. 'Specific' najlepsze dla danego zbioru w kroswalidacji, a 'Default' domyślne z pakietu scikit-learn.

Literatura

[1] P. Probst, B. Bischl, and A.-L. Boulesteix. Tunability: Importance of hyperparameters of machine learning algorithms, 2018.