

Wprowadzenie do sztucznej inteligencji | cw. 4

Damian D'Souza

Dlaczego na jednym zbiorze jest znacznie lepszy wynik niż na drugim?

Dokładność wyników dla obu zbiorów:

	min	śr	std	max
Breast cancer	56,52%	64,49%	3,38%	70,43%
Mushroom	100,00%	100,00%	0,00%	100,00%

Uśredniona macierz pomyłek dla zbioru breast cancer:

		Klasa rzeczywista	
		Brak nawrotu	Nawrót
Klasa predykowana	Brak nawrotu	64,20	24,10
	Nawrót	16,20	10,50

Uśredniona macierz pomyłek dla zbioru mushroom:

		Klasa rzeczywista	
		Jadalny	Niejadalny
Klasa predykowana	Jadalny	1674,07	0,00
	Niejadalny	0,13	1575,80

Wyniki dla zbioru mushroom są znacznie lepsze, ponieważ danych dotyczących grzybów jest znacznie więcej oraz wiele pojedynczych atrybutów pozwala na dokładne przewidywanie klasy.

- Atrybut “odor” jest głównym czynnikiem wpływającym na wybór klasy przez klasyfikator. Drzewo dla zbioru mushroom jest bardzo płytkie, jedynie grzyb bez zapachu nie może być jednoznacznie sklasyfikowany w pierwszym kroku. Nawet jeśli do wytrenowania zostanie podany tylko ten atrybut, wyniki wciąż są bardzo dobre.
- Atrybut “spore-print-color” też ma duży wpływ na klasyfikację grzybów, w drzewie zawierającym wszystkie atrybuty, jest on drugim badanym po “odor”.

Podobnie jak w przypadku poprzedniego atrybutu wyniki po wytrenowaniu tylko na bazie tego są bardzo dobrze jednak gorsze niż dla "odor"

- Są też inne atrybuty, które w dają zadowalające wyniki po ograniczeniu się do tylko do nich np. "cap-color" i "habitat".

Dokładność wyników dla wybranych atrybutów ze zbioru mushroom:

	min	śr	std	max
odor	98,18%	98,56%	0,15%	98,83%
spore-print-color	85,60%	86,69%	0,44%	87,60%
odor & spore-print-color	99,20%	99,43%	0,08%	99,66%
cap-color & habitat	72,95%	74,10%	0,55%	75,20%

Jak widać już sam atrybut "odor" pozwala na bardzo dokładne określenie klasy grzyba, inne atrybuty też dają dobre wyniki jednak nie tak dobre jak pierwszy.

Rozmiar zestawu danych trenujących też ma wpływ na wyniki, ponieważ jeśli jest zbyt mały niektóre wartości atrybutów mogą nie występować w danych trenujących co powoduje brak dopasowań podczas przewidywania klasy i wyboru tej najczęściej występującej. W następnych badaniach losowo wybierana była dana liczba rzędów, które później były dzielone w proporcjach 3:2 na zbiór trenujący i testujący.

l. rzędów	min	śr	std	max
50	35,00%	65,45%	11,49%	85,00%
100	45,00%	63,07%	6,88%	77,50%
200	50,00%	64,69%	4,85%	76,25%
286	55,65%	65,94%	4,93%	76,52%

Dla zbioru breast cancer widać, że zmniejszenie liczby rzędów nieznacznie pogarsza wyniki, ale zwiększa też się rozrzut wyników spowodowany wyborem najczęściej klasy.

l. rzędów	min	śr	std	max
50	55,00%	92,35%	8,05%	100,00%
100	85,00%	96,70%	3,43%	100,00%
200	91,25%	97,72%	1,94%	100,00%
300	95,83%	98,43%	1,25%	100,00%

Dla zbioru mushroom podobnie widać spadek dokładności, oraz wzrost odchylenia standardowego wraz ze spadkiem liczby rzędów.

Na podstawie przeprowadzonych badań można stwierdzić, że klasyfikator ID3 osiąga różne poziomy skuteczności w zależności od charakterystyki analizowanych danych. W przypadku zbioru Mushroom uzyskano bardzo wysoką dokładność, co wynika z tego, że atrybuty takie jak „odor” i „spore-print-color” jednoznacznie determinują przynależność do klas. Nawet przy ograniczeniu liczby atrybutów model nadal osiąga wysokie wyniki, co świadczy o wyraźnej przewidywalności danych. Reguły decyzyjne tworzone na podstawie tych atrybutów są bardzo skuteczne, co pozwala na dokładne klasyfikowanie grzybów jako jadalnych lub niejadalnych.

Z kolei w przypadku zbioru Breast Cancer osiągnięcie równie wysokiej dokładności jest znacznie trudniejsze. W tym zbiorze atrybuty nie różnicują tak jednoznacznie klas, co prowadzi do większego nakładania się ich charakterystyk pomiędzy nimi. Analizy wykazały również, że zmniejszenie liczby próbek w zbiorze powoduje wzrost odchylenia standardowego wyników oraz nieznaczny spadek dokładności. Świadczy to o większej złożoności i trudności w modelowaniu danych Breast Cancer, które wymagają bardziej zaawansowanych metod klasyfikacji.

Podsumowując, klasyfikator ID3 jest bardzo skuteczny w przypadku zbiorów, w których istnieją dominujące atrybuty jednoznacznie różnicujące klasy, co widać na przykładzie zbioru Mushroom. Natomiast w bardziej złożonych zbiorach, takich jak Breast Cancer, pojedyncze drzewo decyzyjne może być niewystarczające.