C2H5OH ID3 - Dokumentacja

Joanna Sokołowska, Rafał Uzarowicz

8 czerwca 2020

1 Podział pracy

- Rafał Uzarowicz: enkapsulacja w klasie ogólnej, dodanie modyfikacji do algorytmu, obsługa ładowania danych, interfejs linii poleceń, testy jednostkowe, dokumentacja(1-4).
- Joanna Sokołowska: podstawowa funkcjonalność algorytmu, testy wydajnościowe, dokumentacja(3-6).

2 Ważne decyzje projektowe i dodatkowe założenia

1. Klasyfikacja z wieloma wartościami klasy odbywa się na podstawie jednego drzewa decyzyjnego, w którego liściach wybierana jest najpopularniejsza klasa w podzbiorze.
2. Zostały zaimplementowane trzy sposoby obsługi danych numerycznych:
   (a) dane dzielone są na dwa podzbiory na podstawie pivota,
   (b) dane dzielone są na zakresy,
   (c) dane są dzielone w klasyczny sposób - każda unikatowa wartość atrybutu to nowa gałąź drzewa.
3. Pivot jest wybierany jako średnia wszystkich wartości danego atrybutu w rozpatrywanym podzbiorze.
4. Liczba zakresów jest równa części całkowitej średniej liczby unikatowych wartości atrybutów.
5. Do algorytmu można podać listę wartości, które zostaną uznane za numeryczne. Jeśli taka lista nie zostanie podana, algorytm sam sprawdzi wszystkie atrybuty i jeśli jakiś będzie zawierał tylko wartości numeryczne, to zostanie uznany za numeryczny. W przypadku podania jakiejkolwiek listy (także pustej) wszystkie pozostałe atrybuty będą traktowane tak, jak podano w punkcie c.
6. Dane nie mogą posiadać pustych wartości w krotce - taka sytuacja jest uznawana za sytuację której dane drzewo nie obejmuje.
7. Jeśli dla danej krotki wartość danego atrybutu nie występuje w wartościach tego atrybutu w drzewie to także jest uznawane, że drzewo nie obejmuje takich wartości.

3 Wykorzystane narzędzia i biblioteki

pandas, numpy, matplotlib, seaborn

4 Opis danych

W testowanych danych znajdują się 1044 krotki i 33 atrybuty, z czego 16 atrybutów numerycznych, 13 binarnych i 4 wyrażone słownie. Z pośród atrybutów numerycznych 11 przyjmuje 5 wartości unikatowych, 1 - 8 wartości, 3 z nich 17 lub 18 wartości, a jeden 34 wartości unikatowe. Atrybuty słowne mają pomiędzy 3 a 5 wartości unikatowych. Część całkowita średniej liczby unikatowych wartości w atrybucie wynosi 5. Żadna z krotek nie zawiera pustych atrybutów. Rozkład wartości w atrybutach klasyfikacji jest następujący:

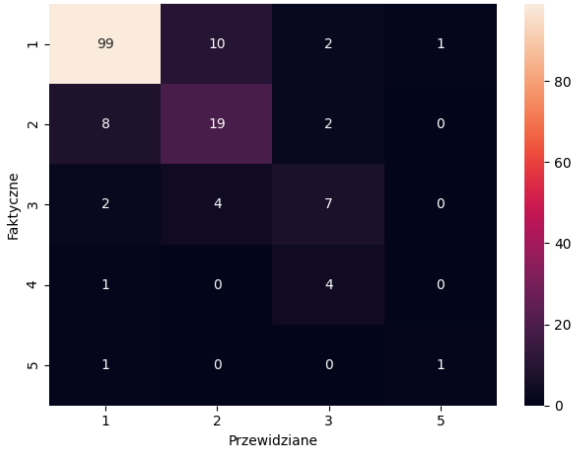
	1	2	3	4	5
Dalc	727	196	69	26	26
Walc	398	235	200	138	73

5 Testowanie algorytmu

5.1 Walidacja krzyżowa na zbiorze *Student alkohol consumption*

Doświadczenie wykonano w celu sprawdzenia dokładności przewidywań dokonywanych przez zaimplementowany algorytm. Zakładamy, że dokładność na zbiorach treningowych będzie znacząco wyższa od tej na zbiorze testowym.

Test przeprowadzono dla obu atrybutów dotyczących konsumpcji alkoholu tj. *Walc* czyli konsumpcja alkoholu w weekendy i *Dalc* czyli dzienna konsumpcja alkoholu. Na potrzeby testu zdecydowano, że wszystkie atrybuty numeryczne będą rozdzielane na dwa z użyciem pivota. Dokonano podziału na 10 podzbiorów i uśredniono wyniki. Otrzymano następujące dokładności: 82,7% dla atrybutu *Dalc* i 68,22% dla atrybutu *Walc*. Na zbiorze treningowym w obu przypadkach otrzymano dokładność 100%.



Rysunek 1: Macierz pomyłek dla klasyfikacji atrybutu *Dalc*

Prawdopodobną przyczyną niedokładności klasyfikacji jest przeuczenie oraz dysproporcja pomiędzy ilością przykładów poszczególnych klas. Dla atrybutu *Dalc* ponad 70% przykładów należy do klasy 1, dla atrybutu *Walc* jest to jedynie 40%, jednak w obu przypadkach klasy 4 i 5 nie posiadają adekwatnej reprezentacji, żeby zostały poprawnie sklasyfikowane.

5.2 Dokładność klasyfikacji w zależności od rozmiaru zbioru treningowego

Eksperyment przeprowadzono, aby zbadać dokładność klasyfikacji dla atrybutu *Dalc* w zależności od rozmiaru zbioru treningowego. Zakładamy, że wraz ze wzrostem części danych przeznaczonych na zbiór treningowy będzie rosła dokładność klasyfikacji na zbiorze testowym.

Podzielono zbiór na część testową i treningową w proporcjach od 0,4 do 0,95 zbioru jako zbiór treningowy, na jego podstawie przeprowadzono klasyfikację, a następnie mierzono dokładność przewidywań algorytmu na pozostałej części zbioru. Dla każdego rozmiaru zbioru treningowego wykonano 10 powtórzeń i uśrednione wyniki zapisano w tabeli.

rozmiar	0,4	0,5	0,6	0,7	0,8	0,9	0,95
dokładność[%]	67,60	70,59	71,44	75,82	76,94	80,61	80,05

Dokładność klasyfikacji rośnie zgodnie z oczekiwaniami, z wyjątkiem klasyfikacji dla zbioru treningowego stanowiącego 95% dostępnych danych. Świadczy to najpewniej o zjawisku przeuczenia - drzewo decyzyjne jest zbyt dobrze dopasowane do danych treningowych, przez co jego generalizacja na inny zbiór jest znacząco słabsza. Najlepsze wyniki klasyfikacji otrzymano dla proporcji zbioru testowego do treningowego 20-80 i 10-90, co wydaje się adekwatnym rozmiarem, gdzie w zbiorze treningowym znajduje się dostatecznie wiele przykładów z wszystkich klas, a zbiór testowy pozostaje wciąż dość obszerny, by adekwatnie weryfikować poprawność klasyfikacji.

5.3 Wpływ zakłóceń na dokładność klasyfikacji

Eksperyment przeprowadzono w celu zbadania wpływu zakłóceń na dokładność klasyfikacji. Spodziewamy się, że większe zakłócenia spowodują mniejszą dokładność klasyfikacji. Dane podzielono na zbiór testowy i treningowy w proporcjach 15-85. Zakłócano odpowiednio zbiór testowy lub treningowy lub oba.Z prawdopodobieństwem równym poziomowi zakłóceń, dla każdej krotki i każdego atrybutu (z wyjątkiem docelowego atrybutu klasyfikacji) zmieniano jego wartość na losowo wybraną spośród innych jego wartości. Następnie zbudowano 2 drzewa - jedno na podstawie niezakłóconego zbioru testowego, a drugie na podstawie zbioru zakłóconego. Dokonano 3 klasyfikacji - dla niezakłóconego zbioru testowego na drzewie zakłóconym i dla zakłóconego zbioru testowego na obu drzewach. Dla każdego poziomu zakłóceń eksperyment powtórzono 10-krotnie i uśrednione wyniki zgromadzono w tabeli.

poziom zakłóceń	czysty testowy, zakłócony treningowy	zakłócony testowy, czysty treningowy	zakłócony testowy i treningowy
0,05	71,47%	67,73%	71,32%
0,15	67,34%	59,05%	64,37%
0,25	62,46%	55,62%	54,93%
0,35	59,55%	53,54%	49,84%
0,45	55,22%	49,89%	45,43%
0,55	53,77%	49,93%	42,12%
0,65	47,86%	48,44%	38,17%
0,75	43,11%	49,11%	36,34%
0,85	43,49%	50,37%	35,60%
0,95	36,90%	52,00%	31,74%

Do przekroczenia pewnego poziomu zakłóceń algorytm radzi sobie z klasyfikacją na spodziewanym poziomie dokładności. Najlepsze wyniki daje klasyfikacja na poziomie zakłóconego

drzewa, zarówno dla przykładów zakłóconych jak i poprawnych. Początkowo ID3 wciąż znajduje znaczące argumenty do wpisywania w węzły drzewa, które dobrze odzwierciedlają faktyczne relacje między danymi, jednak wraz ze wzrostem poziomu zakłóceń dokładność klasyfikacji spada drastycznie. Ten spadek spowodowany jest tym, że pozorne związki między danymi wywołane wprowadzony szumem zaczynają przeważać nad faktycznymi, znaczącymi relacjami.

5.4 Porównanie dokładności klasyfikacji w zależności od wyboru obsługi atrybutów numerycznych

Eksperyment przeprowadzono w celu zbadania wpływu wyboru obsługi arybutów numerycznych na dokładność klasyfikacji. Przeprowadzono 10-krotną walidację krzyżową dla każdego sposobu obsługi arybutów i uśredniono wyniki. Wybrano 3 sposoby obsługi arybutów numerycznych - wszystkie dzielone pivotem, wszystkie dzielone na zakresy i atrybuty o min 8 wartościach dzielone pivotem, a pozostałe traktowane jak napisy.

obsługa atrybutów numerycznych	dokładność
pivot (a)	69,63%
zakresy (b)	23,2%
mieszane (c)	68,72%

Wśród wyników wyraźnie odbiegają te dla zakresów - prawdopodobnie wynika to z tego, że w zbiorze testowym może łatwo trafić się sytuacja, w której w danym zakresie atrybutu nie trafia się żadna wartość, przez co w liściu wpisywana jest klasa nieznana, a późniejsze przewidywania są mniej dokładne. Zastosowanie podejścia mieszanego i rozdzielania pivotem daje podobne wyniki, co świadczy na korzyść rozdzielania pivota w wyniku którego dostajemy prostsze i ogólniejsze drzewo.

6 Podsumowanie

Algorytm ID3 pomimo dobrego dopasowania do zbioru treningowego, osiąga znacznie gorsze wyniki na zbiorze testowym, łatwo ulegając przeuczeniu. Przy nieadekwatnej reprezentacji klas uczy się klasyfikować poprawnie najpopularniejsze z nich, niestety kosztem tych gorzej reprezentowanych. Dość dobrze radzi sobie z wpływem zakłóceń,póki nie osiągną one zbyt wysokiego poziomu, choć generalnie lepiej klasyfikuje zakłócone dane na podstawie zakłóconego drzewa niż na podstawie czystego drzewa. Prawdopodobnie lepiej nadaje się do klasyfikacji danych z mniejszą ilością atrybutów, gdyż wtedy trudniej się przeucza i lepiej odzwierciedla faktyczne związki pomiędzy danymi. Dużą wadą algorytmu ID3, której nie sposób pominąć, jest to, że jest to algorytm zachłanny, przez co budowanie drzewa jest procesem czasochłonnym, podobnie jak wiarygodne testowanie algorytmu z użyciem adekwatnej liczby powtórzeń.