

# [PSZT-U] Magic mushrooms

Piotr Frątczak  
(300207)

Bartosz Świtalski  
(300279)

12 stycznia 2021

## 1 Opis problemu

Implementacja przewidywania jadalności grzyba<sup>1</sup> za pomocą algorytmu ID3.

## 2 Decyzje projektowe

- dane trenujące pochodzą ze [strony](#)

## 3 Cele eksperymentu

Implementacja algorytmu konstruowania drzewa decyzyjnego ID3 z testami binarnymi.

## 4 Użycie

```
/magic-mushrooms$
```

```
cd magic-mushrooms/  
python3 main.py <plik> <indeks> <separator>  
<k> <liczba_uruchomien>
```

### Oznaczenia argumentów

**plik** - nazwa pliku umieszczonego w katalogu **data/**, np. *agaricus-lepiota.data*

**indeks** - indeks pozycji klasyfikatora w liście atrybutów, np. 0

**separator** - separator wartości atrybutów w pliku, np. ,

**k** - wartość  $k$  dla walidacji krzyżowej

**liczba uruchomień** - liczba uruchomień testowych, z których będzie podawany średni wynik

---

<sup>1</sup>rząd pieczarkowców ([agaricales](#))

## Komentarz do użycia

Po uruchomieniu skryptu na wyjściu pojawi się informacja o ustawionych parametrach, tj. liczbie uruchomień, wartość  $k$ , średniej dokładności algorytmu (w zakresie  $[0; 1]$ ) oraz poszczególnych średnich wartościach w macierzy błędów<sup>2</sup>.

## 5 Testowanie

### Analiza testowanego zbioru danych

- liczba przykładów: 8124
- liczba atrybutów: 22
- liczba klas: 2 (edible = e, poisonous = p)
- rozkład klas: edible: 4208 (51.8%), poisonous: 3916 (48.2%)
- liczba brakujących wartości: 2480 (wszystkie oznaczone '?', wszystkie dla atrybutu nr 11 (*stalk-root*))

### Założenia

- w zbiorze danych brak wyróżnienia na podzbiór trenujący oraz testujący
- zastosowano walidację krzyżową ( $k$ –krotną)
- prezentowane wyniki są uśrednieniem z 25 uruchomień programu
- testowaliśmy wpływ wartości parametru  $k$  walidacji krzyżowej na wynik dokładności działania implementowanego algorytmu klasyfikacji
- testowaliśmy klasyfikację bez uwzględnienia najbardziej wartościowych atrybutów (tych, dla których przyrost informacji był największy) na dokładność klasyfikacji
- testowaliśmy wpływ wielkości zbioru danych na dokładność klasyfikacji
- wynik pozytywny oznacza, że grzyb został sklasyfikowany jako jadalny
- wynik negatywny oznacza, że grzyb został sklasyfikowany jako trujący
- wyniki zostały podane w procentach i zostały zaokrąglone do dwóch miejsc po przecinku

---

<sup>2</sup>[wikipedia: Confusion matrix](#)

## 6 Wyniki

### Oznaczenia

TP - średnia liczba prawdziwie pozytywnych wyników (*true positive*)

FN - średnia liczba fałszywie negatywnych wyników (*false negative*)

FP - średnia liczba fałszywie pozytywnych wyników (*false positive*)

TN - średnia liczba prawdziwie negatywnych wyników (*true negative*)

### Wpływ wartości parametru $k$ walidacji krzyżowej

| k  | TP      | FN   | FP   | TN      | dokładność<br>klasyfikacji |
|----|---------|------|------|---------|----------------------------|
| 2  | 4207.68 | 1.64 | 0.32 | 3914.36 | $\approx 99.98\%$          |
| 3  | 4208.0  | 0.32 | 0.0  | 3915.68 | $\approx 100\%$            |
| 5  | 4205.76 | 0.32 | 0.0  | 3913.92 | $\approx 100\%$            |
| 7  | 4205.56 | 0.16 | 0.0  | 3914.28 | $\approx 100\%$            |
| 11 | 4204.8  | 0.08 | 0.0  | 3913.12 | $\approx 100\%$            |
| 17 | 4204.8  | 0.08 | 0.0  | 3913.12 | $\approx 100\%$            |

### Analiza dokładności bez wybranych atrybutów

- testowano dla wartości  $k = 3$  ( $k$ -krotna walidacja).
- 5 najlepszych atrybutów to: *odor* (indeks 5), *spore-print-color* (20), *gill-color*(9), *ring-type* (19) oraz *stalk-surface-above-ring* (12)

Przebadaliśmy przypadki usunięcia kolejno  $n$  *najlepszych* atrybutów. Po zignorowaniu pojedynczego atrybutu sprawdzaliśmy, jaki atrybut został wytypowany na kolejny *najlepszy* i w kolejnym teście dodawaliśmy go do ignorowanych, sprawdzając tym samym kolejny *najlepszy* atrybut itd.

| n brakujących<br>najlepszych atrybutów | TP      | FN   | FP   | TN      | dokładność<br>klasyfikacji |
|--|---------|------|------|---------|----------------------------|
| 1                                      | 4207.72 | 0.92 | 0.28 | 3915.08 | $\approx 99.99\%$          |
| 2                                      | 4207.04 | 1.16 | 0.96 | 3914.84 | $\approx 99.97\%$          |
| 3                                      | 4208.0  | 0.72 | 0.0  | 3915.28 | $\approx 99.99\%$          |
| 4                                      | 4207.8  | 0.32 | 0.2  | 3915.68 | $\approx 99.99\%$          |
| 5                                      | 4207.76 | 0.16 | 0.24 | 3915.84 | $\approx 100\%$            |

## Analiza dokładności wg wielkości zbioru danych

- testowano dla wartości  $k = 3$  ( $k$ -krotna walidacja)
- zbiór zmniejszano o rząd wielkości w każdej iteracji (potęga liczby 2)

| wielkość zbioru | procent zbioru   | TP     | FN   | FP   | TN      | dokładność klasyfikacji |
|-----------------|------------------|--------|------|------|---------|-------------------------|
| 8124            | $\approx 100\%$  | 4208.0 | 0.32 | 0.0  | 3915.68 | $\approx 100\%$         |
| 4062            | $\approx 50\%$   | 3326.0 | 1.0  | 0.0  | 735.0   | $\approx 99.98\%$       |
| 2031            | $\approx 25\%$   | 1791.0 | 1.0  | 0.0  | 239.0   | $\approx 99.95\%$       |
| 1016            | $\approx 13\%$   | 911.12 | 0.0  | 0.0  | 102.88  | $\approx 100\%$         |
| 508             | $\approx 6\%$    | 458.08 | 0.0  | 0.0  | 48.92   | $\approx 100\%$         |
| 254             | $\approx 3\%$    | 223.12 | 0.0  | 0.0  | 28.88   | $\approx 100\%$         |
| 127             | $\approx 1.6\%$  | 107.12 | 0.24 | 0.0  | 18.64   | $\approx 99.81\%$       |
| 64              | $\approx 0.08\%$ | 49.0   | 0.0  | 0.16 | 13.84   | $\approx 99.75\%$       |
| 32              | $\approx 0.4\%$  | 19.88  | 0.08 | 0.76 | 9.28    | $\approx 97.2\%$        |
| 16              | $\approx 0.2\%$  | 9.8    | 0.6  | 1.48 | 3.12    | $\approx 86.13\%$       |

## 7 Wnioski

Dla wartości  $k = 2$  (walidacja krzyżowa) dokładność klasyfikacji algorytmu po uczeniu się na analizowanych zbiorze wynosi  $\approx 99.98\%$ . Przy każdej wartości  $k > 2$  dokładność wzrasta do  $100\%$ , zatem  $k = 3$  zostało wykorzystane do przeprowadzenia dalszych badań dokładności klasyfikacji przy manipulacji innymi parametrami.


Zauważyliśmy, że algorytm działa wyjątkowo dobrze nawet po usunięciu z rozważania kilku najbardziej wartościowych atrybutów (tych o największej potencjalnej zdobyczy informacyjnej). Jest to możliwe, ponieważ podczas normalnego działania algorytmu konstruowane jest drzewo, w którym do klasyfikacji nie są wykorzystywane wszystkie atrybuty, więc przy usunięciu kilku z nich używanych do konstrukcji drzewa binarnego, algorytm zastępuje je dotychczas nieużywanymi i zachowuje swoją dokładność.

Po przeprowadzeniu analizy dokładności przy zmniejszaniu wielkości zbioru danych byliśmy w stanie stwierdzić, że - pomijając wahania losowe - dokładność zmniejsza się o znaczący rząd wielkości przy wielkości ograniczonej do  $\approx 1.6\%$ .

## 8 Podsumowanie

Projekt wprowadzający w tematykę uczenia maszynowego oraz drzew decyzyjnych. Dzięki własnej implementacji algorytmu ID3 poznano istotę konstrukcji drzew decyzyjnych. Zastosowanie testów binarnych pozwoliło na zapoznanie się z przykładową reprezentacją modelu drzewa decyzyjnego.

## 9 Powiązane linki

 [Repozytorium projektowe](#)