Raport końcowy
IREP++

Uczenie się maszyn

Mateusz Szychiewicz

Tematem projektu jest przygotowanie implementacji algorytmu do indukcji reguł **IRep++** oraz przeprowadzenie eksperymentów na zadaniu **Wine Quality**. Dostarczone zostały zbiory danych dotyczące czerwonych i białych wariantów portugalskiego wina „Vinho Verde", zawierające zmienne fizykochemiczne i sensoryczne.

IREP++ jest algorytmem, uczącym się reguł, podobnym do algorytmów RIPPER i IREP. Podobnie do nich, IREP++ generuje dokładne, czytelne dla człowieka reguły z zaszumionych zestawów danych, jednak potrafi on tworzyć zestawy reguł szybciej i często może zawrzeć docelową koncepcję przy użyciu mniejszej liczby reguł i literałów na regułę. Dzięki temu określa opis koncepcji, który jest łatwiejszy do zrozumienia dla ludzi, a jego szybkość pozwala na interaktywny trening przy użyciu bardzo dużych zestawów danych.

1. Dane

Ponieważ algorytm indukcji reguł IRep++ jest jedynie przeznaczony do definiowania wartości dwuklasowych, problemem, do którego rozwiązania został on zastosowanym, było **określenie koloru wina** na podstawie składu chemicznego oraz subiektywnej oceny jakości dokonanej przez ekspertów.

Zmienne wejściowe (na podstawie testów fizykochemicznych):
1 - stała kwasowość
2 - lotna kwasowość
3 - kwas cytrynowy
4 - cukier resztkowy
5 - chlorki
6 - wolny dwutlenek siarki
7 - całkowity dwutlenek siarki
8 - gęstość
9 - pH
10 - siarczany
11 - alkohol
12 - jakość (wynik od 0 do 10)

Zmienna wyjściowa (na podstawie danych sensorycznych):
Kolor - biały lub czerwony

2. Implementacja

Napisany w języku Python program pozwala na **naukę reguł** oraz na ich **walidację**.
Podając pliki zawierające atrybuty oraz ich wartości, tworzy model zawierający zbiór reguł, który zapisywany jest w ustalonym przez użytkownika folderze. W celu wykonania testu należy podać ścieżkę do wcześniej stworzonego modelu, za pomocą którego definiowany będzie atrybut.

Wymagane są następujące **argumenty**:

-e: tryb wykonania (learn | classify)
-a: ścieżka do pliku atrybutów
-c: nazwa definiowanego atrybutu w pliku atrybutów
-t: ścieżka do pliku treningowego / testowego
-m: ścieżka do pliku modelu (wyniki odczytywalne maszynowo)
-o: ścieżka do pliku wyjściowego (wyniki czytelne dla człowieka)

Przykładowa komenda służąca do nauki:

```
python irep.py -e learn -a "./attr/wine.txt" -c color -t
"./train/wine-train.txt" -m "./results/wine-model.dat" -o
"./results/wine-model.txt"
```

Przykładowa komenda służąca do testu:

```
python irep.py -e classify -a "./attr/wine.txt" -c color -t
"./test/wine-test.txt" -m "./results/wine-model.dat" -o
"./results/wine-model-test.txt"
```

Gdzie przykładowe pliki mają następujący format:

Plik atrybutów:
`wine.txt`

```
fixed_acidity float
volatile_acidity float
citric_acid float
residual_sugar float
chlorides float
free_sulfur_dioxide float
total_sulfur_dioxide float
density float
pH float
sulphates float
alcohol float
quality int
```

```
color white red
```

Plik treningowy/testowy:
`wine-train/wine-test.txt`

```
7.7 0.54 0.26 1.9 0.089 23 147 0.99636 3.26 0.59 9.7 5 red
6.9 0.74 0.03 2.3 0.054 7 16 0.99508 3.45 0.63 11.5 6 red
6.6 0.895 0.04 2.3 0.068 7 13 0.99582 3.53 0.58 10.8 6 red
7 0.27 0.36 20.7 0.045 45 170 1.001 3 0.45 8.8 6 white
6.3 0.3 0.34 1.6 0.049 14 132 0.994 3.3 0.49 9.5 6 white
8.1 0.28 0.4 6.9 0.05 30 97 0.9951 3.26 0.44 10.1 6 white
```

Pliki zawierające model:
`wine-model.dat`

```
{'red': [{'total_sulfur_dioxide': ('<=', 65.0)}]}
```

`wine-model.txt`

```
IF total_sulfur_dioxide <= 65.0 THEN red
ELSE white
```

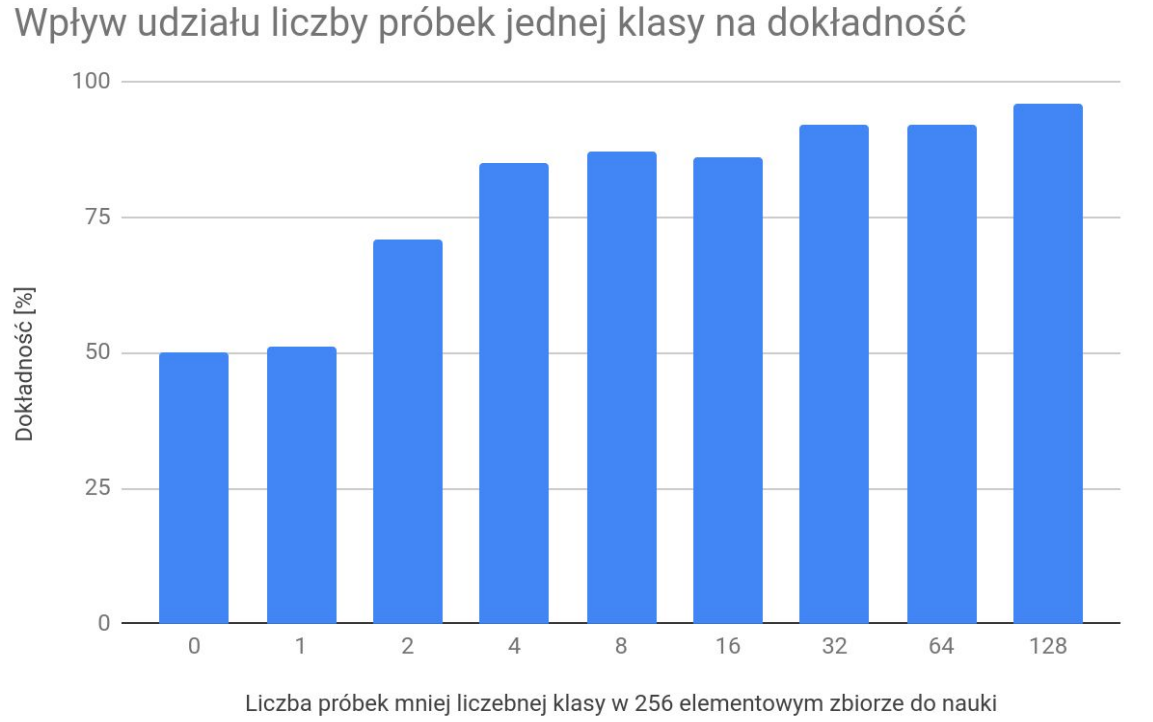
Plik zawierający wyniki testu:
`wine-model-test.dat`

```
Class         Cases        Classified

white        200          192
red          100          90

Accuracy: 94.0%
```

3. Testy

Pierwszy test miał na celu określenie wpływu liczby próbek w zbiorze treningowym na otrzymaną dokładność otrzymanego zbioru reguł. W tym celu wielkość zbioru była stopniowo zwiększana, a dla każdej wielkości wykonanych zostało 10 testów, których otrzymany rezultat został uśredniony. Spodziewanym rezultatem była rosnąca dokładność wraz ze wzrostem liczby próbek, jednak przeprowadzone na zadaniu Wine Quality testy wykazały, że optymalnym rozmiarem zbioru treningowego jest 256 próbek.

Celem kolejnego testu było określenie wpływu udziału próbek jednej klasy na dokładność otrzymanego zbioru reguł. Dla optymalnej wielkości zbioru treningowego, otrzymanego w poprzednich badaniach, przeprowadzone zostały testy, w których udział próbek jednej klasy zmieniany był od 0% do 50%. Podobnie jak poprzednio dla każdej wartości udziału, przeprowadzono 10 prób, których wyniki zostały uśrednione. Tym razem jednak, zgodnie z przewidywaniami, największa dokładność została uzyskana dla równego podziału próbek między dwie klasy.



Ostatnim wykonanym testem był test mający na celu ustalenie przydatności poszczególnych atrybutów dostępnych w zadaniu Wine Quality na dokładność określenia koloru wina.
W tym celu osobno dla każdego z atrybutów przeprowadzona została seria testów, a otrzymana dokładność uśredniona.

Trzy, mające największy wpływ na dokładność ustalenia koloru wina, parametry to kolejno:

- zawartość chlorków
- gęstość
- zawartość alkoholu

Jak widać na wykresie połowa atrybutów ma bliski zerowemu wpływ na otrzymywany wynik. Co ciekawe jakość jako jedyna nie ma żadnego wpływu na kolor wina.