AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe" – laboratorium

Laboratorium 3

Klasyfikacja

Indukcja drzew decyzyjnych za pomocą C4.5 (C5.0)

data aktualizacji: 16.04.2022

Cel ćwiczenia

Celem ćwiczenia laboratoryjnego jest zapoznanie się z klasycznym algorytmem C4.5 i jego nowszą wersją (C5.0).

Celem pośrednim ćwiczenia jest zapoznanie się z platformą R, językiem skryptowym oraz dokumentacją niezbędną do realizacji zadania.

Do realizacji zdania użyte powinny być pakiety realizujące programowo algorytmy C4.5 i C5.0, opcje wizualizacji drzew, ocena jakości klasyfikacji (tabele, miary jakości Acc, Fsc) oraz wizualizacji danych w R.

Wprowadzenie

W zadaniu badane będą dwa algorytmy – C4.5 i jego „następca" C5.0. Oba algorytmy działają w oparciu o miarę zysku informacyjnego (*infoGain* opartego na mierze entropii) i na tej podstawie zachłannie dzielą dane budując finalne drzewo decyzyjne. Każdy z tych

algorytmów ma zestaw parametrów, które mogą bardzo krytycznie wpłynąć na skuteczność klasyfikacji wynikowego drzewa decyzyjnego.

Sugerowane narzędzia

Narzędzia sugerowane do realizacji tego zadania są ogólnodostępne, w większości na licencji *open source*. Do realizacji zdania potrzebne będzie zainstalowanie środowiska R oraz Rstudio (IDE dedykowane do języka R)

Pakiety, które mogą się przydać do realizacji zdania:

- pakiet **caret** (Classification And REgression Training)
  https://cran.r-project.org/web/packages/caret/caret.pdf
- **RWeka** – https://cran.r-project.org/web/packages/RWeka/RWeka.pdf
  Podpowiedź: **C4.5** w RWeka to ~~J23~~ J48
- **C5.0** - https://cran.r-project.org/web/packages/C50/C50.pdf
- Pakiet **ggplot2** do wizualizacji danych –
  dokumentacja: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
  czytelniejsze wprowadzenie: https://www.r-graph-gallery.com/ggplot2-package.html
- też ew. pakiet **mlr** (machine learning in R)
  dokumentacja: https://cran.r-project.org/web/packages/mlr/mlr.pdf
  opis: https://mlr.mlr-org.com/
- **Rmarkdown** – narzędzie do generowania raportów w R (wbudowane w RStudio)
  dokumentacja: https://rmarkdown.rstudio.com

Przebieg ćwiczenia

1. Instalacja platformy R (https://www.r-project.org/) oraz Rstudio (https://www.rstudio.com/products/rstudio/download/) wraz z pakietami niezbędnymi do realizacji zdania.
2. Zaczytanie danych zbiorów IRIS, WINE, GLASS
3. Uruchomienie algorytmu C4.5 dla IRIS dla domyślnych parametrów
4. Wizualizacja drzewa (podpowiedź: *plot*) i analiza jakości klasyfikacji wynikowego drzewa decyzyjnego
5. Wstępne strojenie algorytmów dla zbiorów; dla C4.5 sugerowane 3 parametry:

- C4.5 – *confidence factor*
- C4.5 – *minimum number of instances*
- C4.5 – *pruning*

6. Uruchomienie algorytmu C5.0 dla domyślnej konfiguracji
7. Wstępne strojenie algorytmów dla zbiorów; dla C5.0 sugerowane 4 parametry:
   - C5.0 – *confidence factor*
   - C5.0 – *minCases*
   - C5.0 – *winnow*
   - C5.0 – *noGlobalPruning*
8. Użycie walidacji krzyżowej (też: stratyfikowanej)

Podpowiedź – mały *tutorial* dot. walidacji krzyżowej w R http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/ Stratyfikowana walidacja krzyżowa: https://rdrr.io/cran/SurvRank/man/crossvalFolds.html Caret: https://stackoverflow.com/questions/35907477/caret-package-stratified-cross-validation-in-train-function

9. Użycie wykresów z pakietów do wizualizacji danych (np. ggplot2)
10. Podsumowanie wyników. To jest miejsce na tabelki, wykresy, wnioski – wybieramy prezentowane dane/zestawienia, nie dajemy wszystkich wyników.

Uwaga! Przy tym zadaniu nie używamy *boostingu* – ten mechanizm będzie badany przy okazji jednego z następnych zadań laboratoryjnych.

Punktacja

Przy realizacji zadania student może otrzymać **max 10 punktów** wedle poniższej tabeli.

2	Zapoznanie się z platformą R – wczytanie danych, uruchomienie klasyfikatora, wizualizacja wyników
2	Zbadanie algorytmu C4.5 – jak 3 wybrane parametry wpływają na skuteczność klasyfikacji 3 zbiorów

2	Zbadanie algorytmu C5.0 – jak 3 wybrane parametry wpływają na skuteczność klasyfikacji 3 zbiorów
2	Zbadanie, jak wyniki użycia walidacji stratyfikowanej różnią się od „zwykłej” walidacji krzyżowej
2	Porównanie wyników działania (tabelki/wykresy) obu algorytmów. Analiza (i narysowanie) drzew decyzyjnych dla 2 wybranych zbiorów.

Przy realizacji tego zadania wystarczy prosty raport PDF. Przy wynikach badań należy dać komentarz, podać wnioski i podsumowanie.

Pytania pomocnicze

1. Co jest modelem klasyfikacji w C4.5 (C5.0)?
2. Co znajduje się w liściach drzewa?
3. Czy przycinanie drzewa (*pruning*) jest potrzebne?
4. Czy drzewo może być za „duże” lub za „małe”?
5. Dlaczego typ/rozmiar walidacji krzyżowej może mieć duży wpływ na skuteczność modelu?
6. Czy C4.5 (C5.0) potrzebuje normalizacji/standaryzacji/dyskretyzacji danych?
7. Czy model można przeuczyć?
8. Na podstawie działania C4.5 (C5.0), wyników i dokumentacji R (parametrów) – na czym polega przewaga C5.0?

Literatura

1. Wykłady do przedmiotu autorstwa prof. H. Kwaśnickiej
2. Cichosz P. ”Systemy uczące się”, WNT Warszawa
3. „Notes for professionals”, http://books.goalkicker.com/RBook/
4. P. Biecek, http://www.biecek.pl/R/ → https://www.ibuk.pl/fiszka/39524/analiza-danych-z-programem-r-modele-liniowe-z-efektami-stalymi-losowymi-i-mieszanymi.html
5. Pełna dokumentacja R https://www.rdocumentation.org/

6. Zasoby Internetu: uczenie maszynowe (machine learning), data mining, R, klasyfikacja, drzewa decyzyjne, indukcja drzew decyzyjnych, C4.5, C5.0, pruning (przycinanie drzewa), generalizacja