Porównanie algorytmów imputacji w języku R

Marceli Korbin, Patryk Wrona, Mikołaj Jakubowski

Plan prezentacji

I. Czym jest imputacja?
II. Co chcemy zbadać? Co chcemy uzyskać?
III. Jakie zbiory nam do tego posłużą?
IV. Które metody przetestujemy?
V. Które modele uczenia maszynowego wykorzystujemy?
VI. Jak przeprowadzamy testy?
VII. Wstępne rezultaty
VIII. Co dalej?

Definicja imputacji

Imputacja - sztuczne zastąpienie braków danych. Stosuje się ją ponieważ wiele modeli, jak na przykład modele uczenia maszynowego, nie akceptuje braków danych. Istnieje wiele różnych technik imputacji danych.

W przypadku R braki danych reprezentowane są jako **NA**.

Co chcemy zbadać/uzyskać?

1. Jak szybko działają poszczególne algorytmy, w zależności od ilości brakujących danych.
2. Jaki wpływ na wyniki predykcyjne mają poszczególne algorytmy imputacji.
3. W jakich sytuacjach lepiej korzystać z prostszych i szybszych, a kiedy z bardziej skomplikowanych i wolniejszych.

Jakie zbiory?

Wykorzystujemy 8 zbiorów ze strony OpenML
https://www.openml.org/search?type=data

Są to zbiory danych o poniższych indeksach:

4, 27, 29, 38, 55, 56, 188, 944

Małe zbiory

labor(4) - Wymiary: 57x17 Braki danych: 326
colic(27) - Wymiary: 368x23 Braki danych: 1927
hepatitis(55) - Wymiary: 155x20 Braki danych: 167
vote(56) - Wymiary: 435x17 Braki danych: 392
echoMonths(944) - Wymiary: 130x10 Braki danych: 97

Duże zbiory

credit-approval(29) - Wymiary: 690x16 Braki danych: 67
sick(38) - Wymiary: 3772x30 Braki danych: 6064
eucalyptus(188) - Wymiary: 736x20 Braki danych: 448

Testowane metody imputacji

- Metody naiwne:
  - Usuwanie kolumn
  - Uzupełnianie średnią/medianą/modą
- Metody zaawansowane:
  - IRMI z pakietu VIM
  - missForest
  - mice

Wykorzystane modele

- XGBoost
- Random Forest
- Logistic Regression
- ...

Jak przeprowadzamy testy?

Na najbliższe działania **stworzyliśmy specjalną funkcję**. Przyjmuje ona w argumentach m.in. zbiór danych, listę zaimplementowanych algorytmów imputacji i listę algorytmów uczenia maszynowego.

Funkcja **mierzy średni czas wykonania każdej imputacji** na zbiorze (z liczby podanych iteracji), a następnie mierzy na uzupełnionym zbiorze **skuteczność algorytmów**. Na następnym slajdzie podajemy stosowane przez nas metryki.

Wynikiem funkcji jest lista macierzy z **czasami imputacji i miarami skuteczności**.

Metryki do porównywania wyników

- Accuracy
- Precision
- Recall
- F1

Wstępne rezultaty

Do tej pory udało nam się dojść do następujących wniosków:

- skomplikowane, czasochłonne techniki imputacji dają lepsze wyniki (niestety na ogromnych zbiorach danych są one niewskazane)

- banalne i szybkie techniki imputacji (np. usuwanie kolumn, uzupełnianie średnią/medianą) dają nieco gorsze wyniki (polecamy je używać na dużych zbiorach danych z małą ilością braków danych)

Co dalej?

- dobranie 4. modelu machine learning
- zrobienie rankingu technik imputacji - bezsensowne techniki imputacji powinny być na ostatnich miejscach (np. usuwające wszystkie kolumny)
- Przetestowanie wszystkich modeli dla każdego ze zbiorów danych
- Wnikliwa analiza czasów i wyników dla poszczególnych algorytmów.