Zestawienie imputacji w R

Marceli Korbin, Patryk Wrona, Mikołaj Jakubowski

Czym jest imputacja?

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**.

Wykorzystane metody imputacji

- IRMI z pakietu VIM
- missForest
- hotdeck z pakietu mice
- kNN
- średnia/mediana/dominanta (jako bazowa metoda podstawowa)

Eksperymenty przeprowadzane na zbiorach z OpenML różnej wielkości

Cele badania

1. Czy imputacja ma znaczący wpływ na wyniki predykcyjne?
2. Czy czasy działania poszczególnych algorytmów, znacząco różnią się między sobą?

MAŁE ZBIORY: 4, 27, 55, 56, 944

Wpływ na wyniki predykcyjne

Acc

	logreg	naiveBayes	binomial	ranger
IRMI	4	3	4	4
missForest	2	1	2	1
hotdeck	3	4	3	3
kNN	5	2	5	4
mean/median/dom	1	5	1	1

F1

	logreg	naiveBayes	binomial	ranger
IRMI	4	2	4	3
missForest	2	1	2	1
hotdeck	3	4	3	3
kNN	5	3	5	5
mean/median/dom	1	4	1	2

DUŻE ZBIORY: 29, 88

Wpływ na wyniki predykcyjne

Acc

	logreg	naiveBayes	binomial	ranger
IRMI	3	1	3	1
missForest	4	2	4	2
hotdeck	5	4	5	2
kNN	1	5	1	2
mean/median/dom	2	3	2	5

F1

	logreg	naiveBayes	binomial	ranger
IRMI	3	1	3	1
missForest	3	1	3	2
hotdeck	5	5	5	5
kNN	1	4	1	3
mean/median/dom	2	3	2	4

WSZYSTKIE ZBIORY

Wpływ na wyniki predykcyjne

Acc

	logreg	naiveBayes	binomial	ranger
IRMI	5	2	5	1
missForest	2	1	2	1
hotdeck	4	4	4	3
kNN	3	3	3	5
mean/median/dom	1	5	1	3

F1

	logreg	naiveBayes	binomial	ranger
IRMI	3	2	3	2
missForest	2	1	2	1
hotdeck	5	5	5	5
kNN	3	3	3	4
mean/median/dom	1	4	1	3

Szybkość działania

mean/med/dom 0 - 17 s IRMI

Dziękujemy za uwagę!