Project A

Mathematical Underpinnings of Machine Learning

Grzegorz Zakrzewski, 313555

Feature selection methods

Method	Stopping rule
Joint Mutual Information	$\lvert S\rvert < 2 \cdot \sqrt{n\_features}$
Conditional Mutual Info Maximisation	$Crit_i > 0.9 \cdot \frac{Crit_{i-1} + Crit_{i-2}}{2}$
Random Forest Feature Importance	$\lvert S\rvert < 2 \cdot \sqrt{n\_features}$
	$FI_i > \frac{1}{2 \cdot n\_features}$
Sequential Feature Selection kNN (k=3)	$\lvert S\rvert < 2 \cdot \sqrt{n\_features}$

Artificial datasets

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\sum_{i=1}^{5} X_i > q_{0.50}$

Table 1: *irrelevant* artificial dataset

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$\mathcal{N}_{0,1}$	$X_1 \sim_{0.9}$	$X_2 \sim_{0.8}$	$X_3 \sim_{0.7}$	$X_4 \sim_{0.6}$	$X_1 \sim_{0.5}$	$\sum_{i=1}^{5} X_i > q_{0.50}$

Table 2: *correlated* artificial dataset

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$B_{0.5}$	$XOR\,(X_1, X_2)$

Table 3: *XOR* artificial dataset

Real-world data examples

Dataset	#rows	#columns	#classes
Abalone	4,177	8	2
Mushroom	58,598	12	2
Students dropout	4,424	36	3

Results assessment

Goal	Method
Effectiveness	accuracy of kNN($k = 3$) with 4-fold CV
Correctness	Bolón-Canedo *Success Rate*
Stability	Kuncheva consistency index

Bolón-Canedo Success Rate

$$Suc. = \left[\frac{R_s}{R_t} - \alpha\frac{I_s}{I_t}\right]$$

- R_s - the number of relevant features selected
- R_t - the total number of relevant features
- I_s - the number of irrelevant features selected
- I_t - the total number of irrelevant features
- $\alpha = \{\frac{1}{2}, \frac{R_t}{I_t}\}$

Kuncheva consistency index

$$C(A, B) = \frac{rn - k^2}{k(n - k)}$$

- the consistency for two subsets $A, B \subset X$
- $|A| = |B| = k$
- $r = |A \cap B|$
- $0 < k < |X| = n$
- $C(A, B) \in [-1, 1]$

Artificial datasets - Success Rate

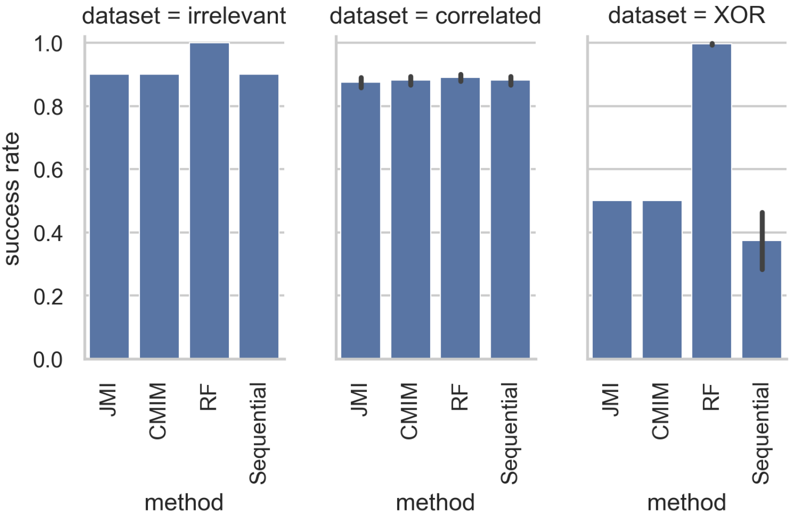

Figure 1: Success Rate computed on artificial datasets.

Artificial datasets - accuracy

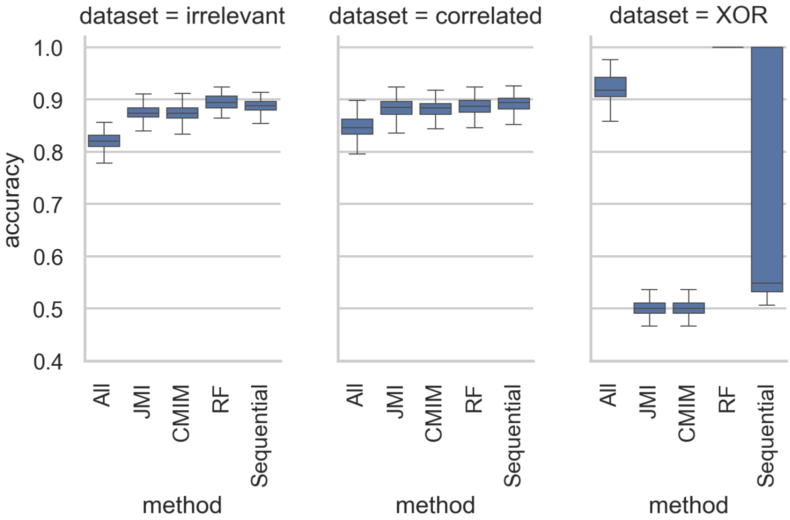

Figure 2: Generic classifier accuracy computed on artificial datasets.

Real-world data examples - stability

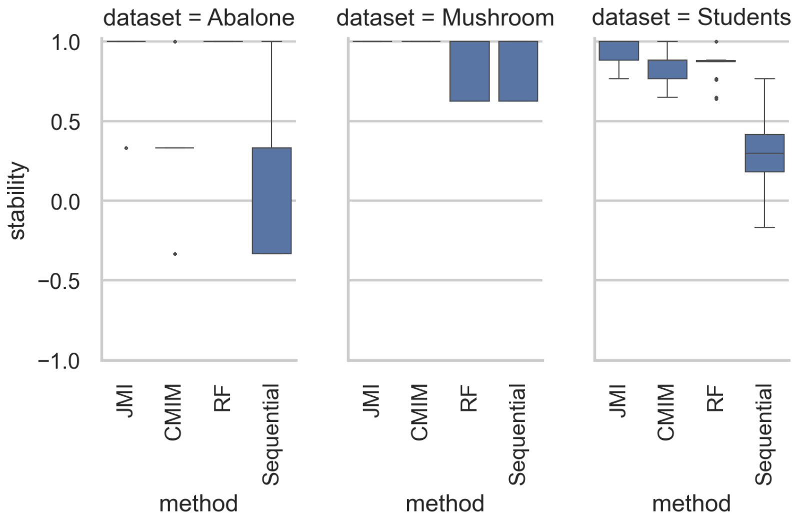

Figure 3: Stability computed on real-world data examples.

Real-world data examples - accuracy

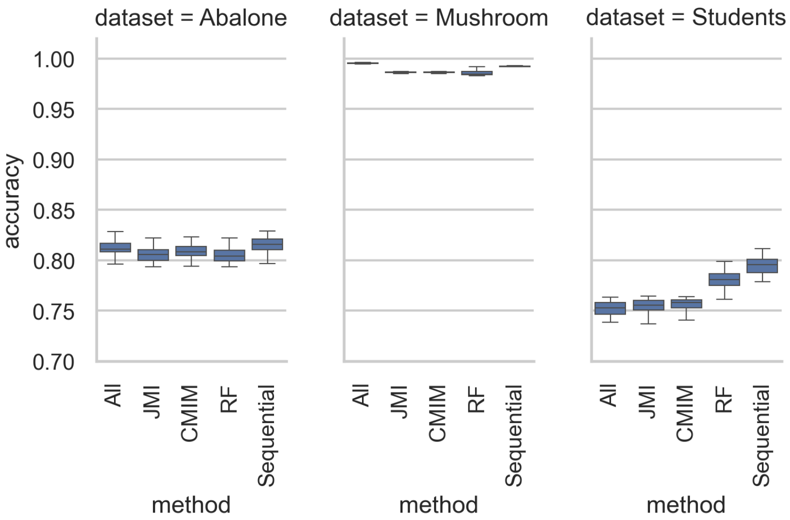

Figure 4: Generic classifier accuracy computed on real-world data examples.

Conclusions

- The project was performed according to the instructions.
- All feature selection methods are working.
- The best method seems to be Random Forest feature importance.
- A much more extended study should be conducted to fully compare their performance.