Feature Selection

Izabela Telejko, Grzegorz Zbrzeżny

Methods Used

Mutual Information based methods:

- Joint Mutual Information Maximisation (JMIM)
- Interaction Gain Feature Selection (IGFS)
- Conditional Mutual Information Maximization (CMIM)

Other methods:

- Sequential L1-based Feature Selection
- Feature Selection based on AIC and BIC

Joint Mutual Information Maximisation

First step:

$$X_{JMIM} = \underset{X_i \in F}{argmax}\ I(X_i; Y)$$

Every next step:

$$X_{JMIM} = \underset{X_i \in F/S}{argmax}[\underset{X_s \in S}{min}\ I(X_i, X_s; Y)]$$

Interaction Gain Feature Selection

$$X_{IGFS} = \underset{X_i \in F/S}{argmax}[I(X_i;Y) + \frac{1}{|S|}\sum_{X_s \in S} I(X_i;X_s;Y)]$$

Conditional Mutual Information Maximization

$$X_{CMIM} = \underset{X_i \in F/S}{argmax}\{I(X_i;Y) - \underset{X_s \in S}{max}[I(X_i;X_s) - I(X_i;X_s|Y)]\}$$

Stopping Rule

$$score(X_i) = NMI(X_i; Y) - \frac{1}{|S|} \cdot \sum_{X_s \in S} NMI(X_i; X_s)$$

where:

- X_i – candidate feature in step i,
- $H(X)$ – entropy of the X variable,
- $NMI(X;Y) = \frac{I(X;Y)}{min[H(X),H(Y)]}$ – normalized mutual information.

Datasets Overview

Dataset	obs.	feat.	relevant feat.
aids	2139	22	-
cancer	569	310	-
divorce	170	54	-
gait	47	321	-
lol	21515	86	-
generated_0	1000	35	5
generated_1	1000	10	5
generated_2	1000	15	5
xor	1000	13	3

Table 1: Datasets basic information after preprocessing.

Artificial Datasets

$$X_1 \sim \mathcal{N}(0,1) \quad (\text{size: } n \times n_{\text{rel}})$$
$$X_2 \sim \mathcal{N}(0,1) \quad (\text{size: } n \times n_{\text{irrel}})$$
$$X_3 = X_1 + \mathcal{N}(0,0.1) \quad (\text{size: } n \times n_{\text{rel}})$$
$$X_4 = \text{interactions}(X_1) \quad (\text{size: } n \times \frac{n_{\text{rel}} \times (n_{\text{rel}} - 1)}{2})$$
$$X = [X_1, X_i], \quad i \in \{2,3,4\}$$
$$Y_{\text{num}} = X_1 \times [\beta_1, \beta_2, \ldots, \beta_{n_{\text{rel}}}]^T + \beta_0$$
$$Y = \text{qcut}(Y_{\text{num}}, n_{\text{classes}}),$$

Artificial Datasets

XOR dataset:

- Three relevant features
- Target variable created in two steps:
  - XOR of first two relevant features
  - XOR of output from preceding step and third relevant feature

Results Evaluation - Artificial Data

Dataset	# Selected	BIC	AIC	CMIM	JMIM	IGFS	L1
generated_0	Rel. Feat.	5	5	5	5	5	5
	Irrel. Feat.	0	6	0	0	0	3
generated_1	Rel. Feat.	5	5	5	4	5	5
	Irrel. Feat.	2	2	0	0	0	0
generated_2	Rel. Feat.	5	5	5	5	5	5
	Irrel. Feat.	0	2	0	0	0	2
xor	Rel. Feat.	0	0	1	0	0	0
	Irrel. Feat.	1	3	0	1	1	1

Table 2: Number of selected relevant and irrelevant features for synthetic datasets.

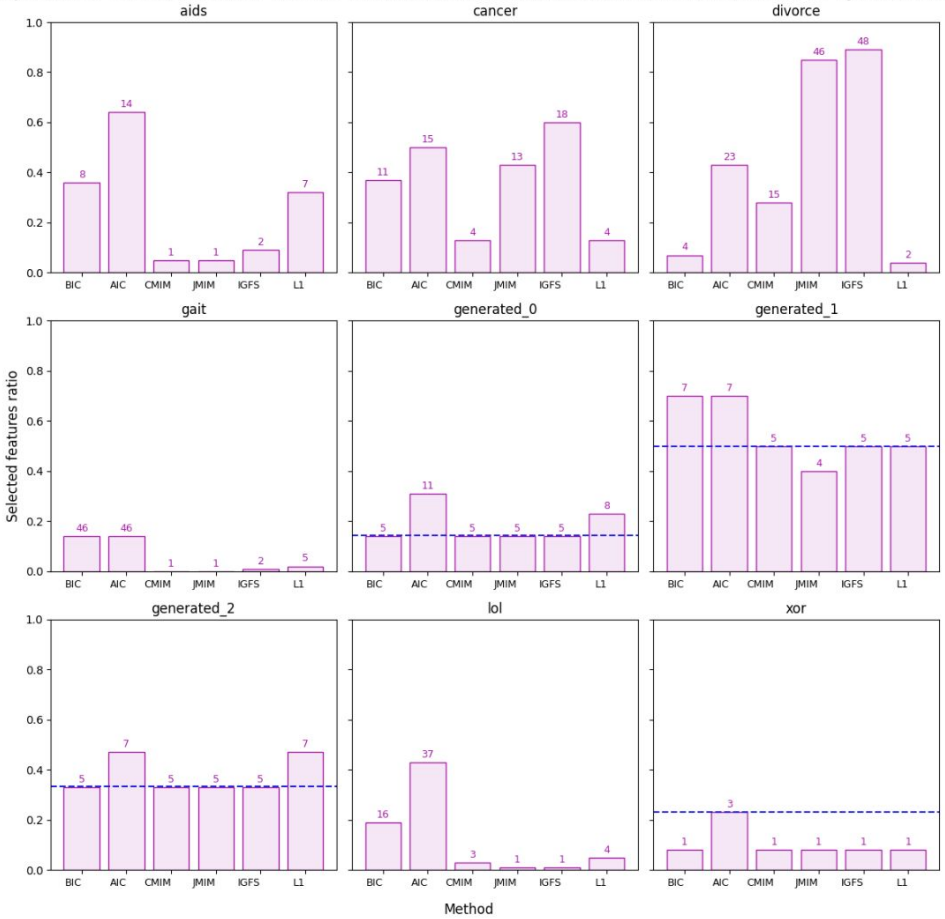

Figure 1: Number of selected features for each method and dataset.

K Nearest Neighbours

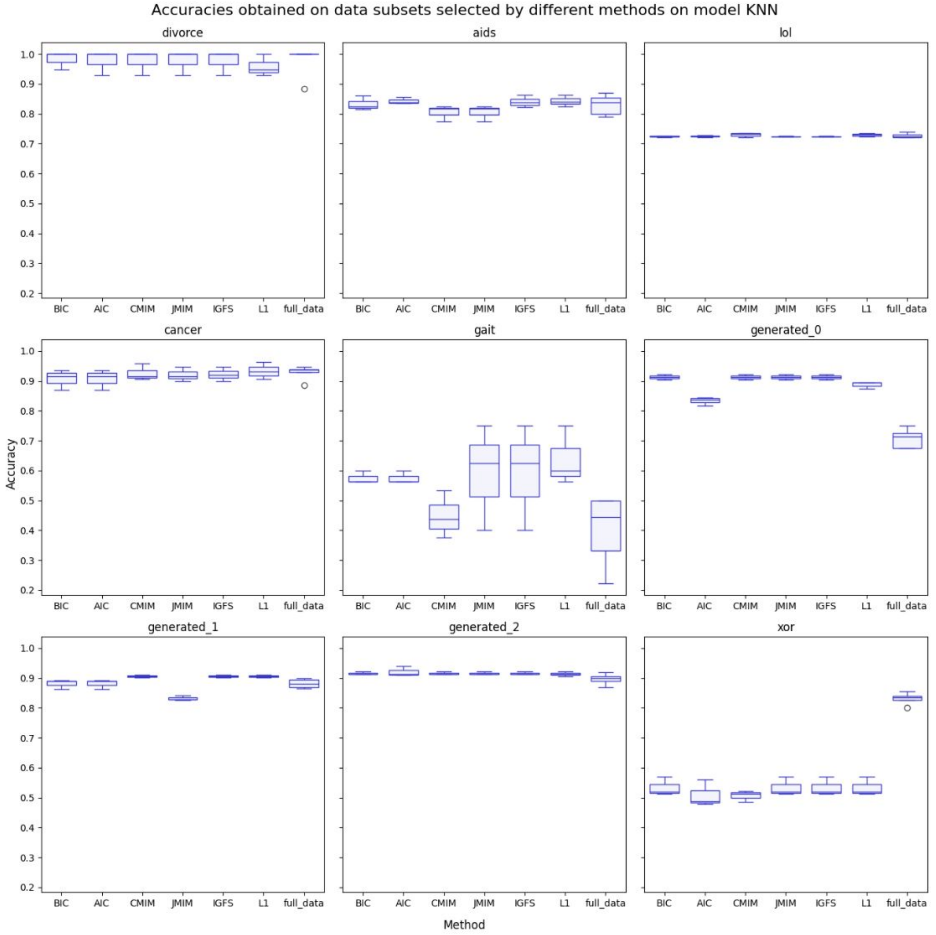

Figure 2: Accuracy scores for KNN models.

Random Forest Classifier

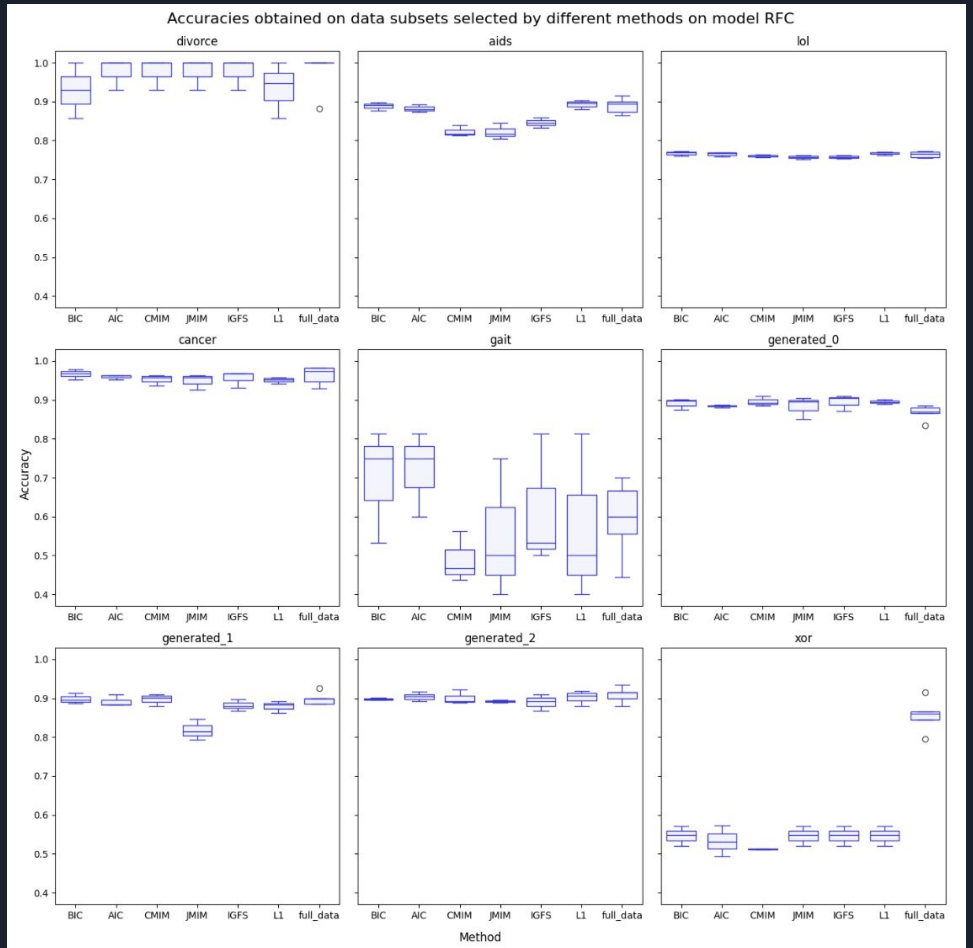

Figure 3: Accuracy scores for RFC models.

Support Vector Classifier

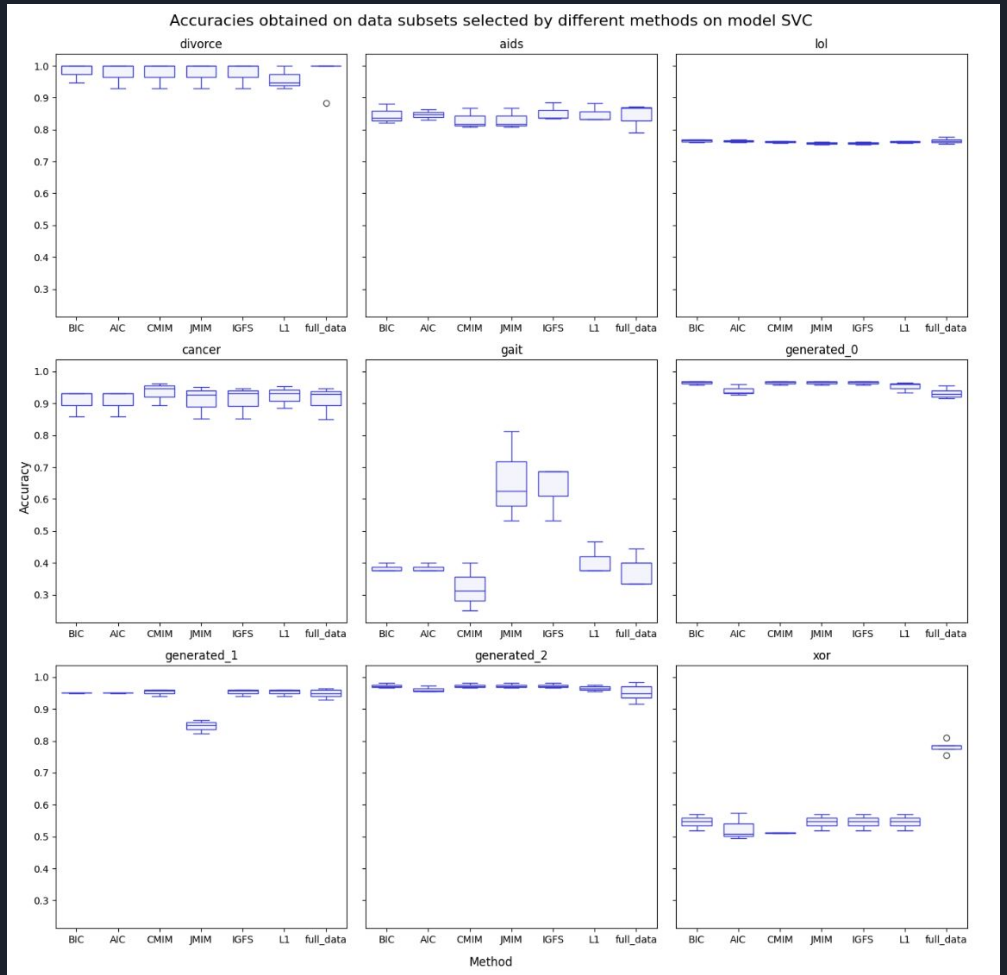

Figure 4: Accuracy scores for SVC models.

Conclusions

- For real-world datasets, models trained on features selected by the MI-based methods achieved slightly higher or very similar scores compared to other referential methods and training on the full dataset,
- For the generated datasets the MI methods accurately detected relevant features with minimal errors,
- Among the MI methods, CMIM generally performed slightly worse than the other two methods,
- Different models achieved varying results on the same sets of selected features.

Thank you for your attention!