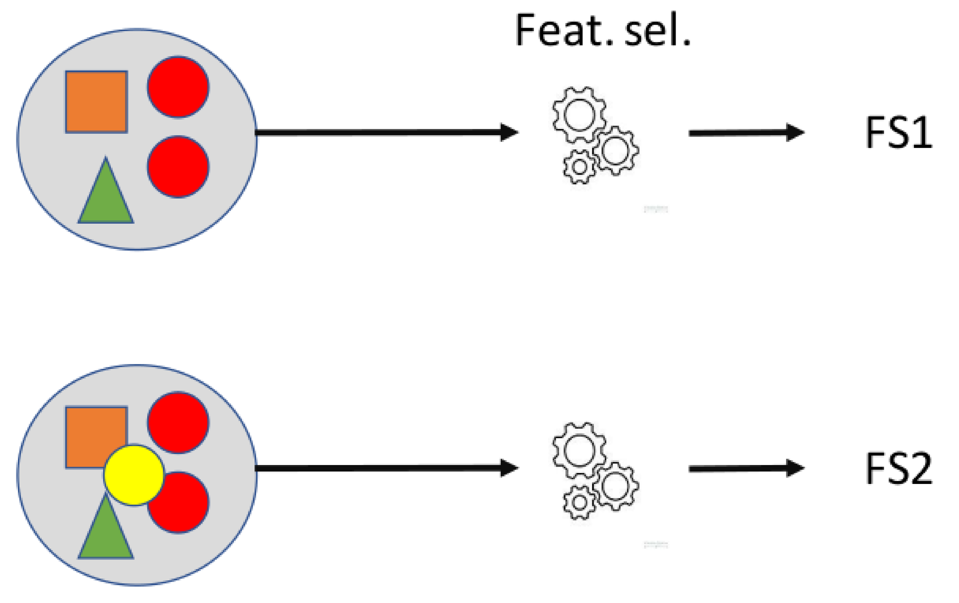
Explicit Control of Feature Relevance and Selection Stability Through Pareto Optimality

Victor Hamer and Pierre Dupont

UCLouvain - ICTEAM/INGI/Machine Learning Group, Louvain-la-Neuve, Belgium.

Introduction

- **Feature selection** (FS) is the act of selecting a small and relevant subset of input features, generally to be included in a predictive model.
  - Reduces overfitting ⇒ improves prediction performance.
  - Learns fast, compact and **easy-to-interpret** models.
- **Selection instability**: selected feature subsets may change drastically after marginal changes in the data.



$$|FS1 \cap FS2| \approx 0 \Rightarrow \text{stab} \downarrow$$

  - Features can be analyzed by **experts** to gain domain knowledge.
  - Instability reduces the interpretability of the predictive models.
  - And the trust of domain **experts** towards the selected features.

State of the literature

- Increasing stability
  - Ensemble feature selection : selects features that are selected the most accross different selection runs.
  - Instance weighting : weights training instances according to their importance to feature evaluation.
  - Model selection: takes stability into account in the fitting of the meta-parameters.

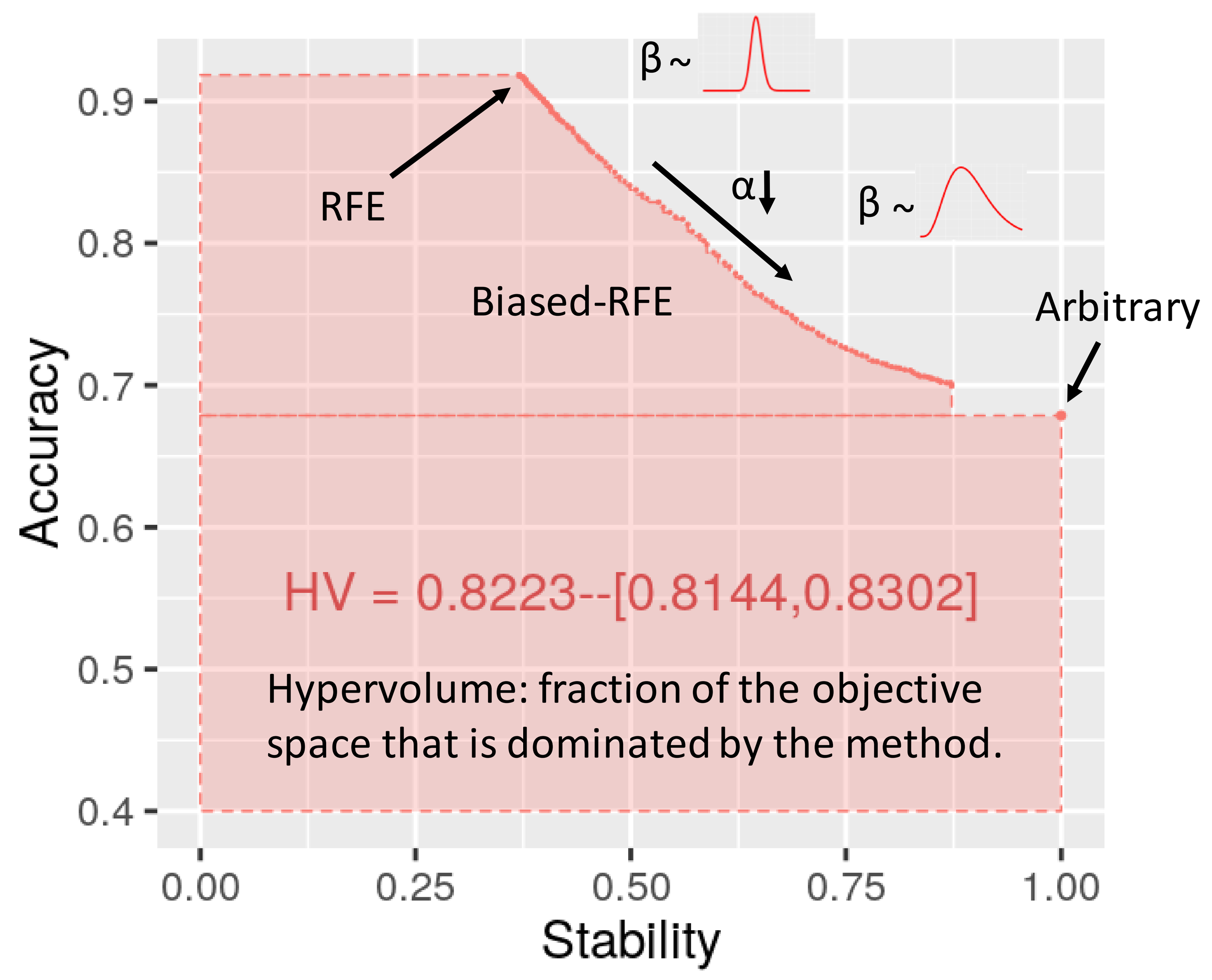
  ⇒ No fine control of the accuracy-stability trade-off.
- Stability measure [1]

$$\phi = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} p_f(1-p_f)}{\frac{k}{d} * (1 - \frac{k}{d})} \quad \begin{cases} d: \text{number of input features} \\ k: \text{mean number of selected features} \\ p_f: \text{feature } f \text{ selection frequency} \end{cases}$$

Biased Logistic RFE

$$L = \sum_{i=1}^{n} \log(1 + e^{-y_i * (\mathbf{w} * \mathbf{x}_i)}) + \lambda \beta ||\mathbf{w}||_2$$

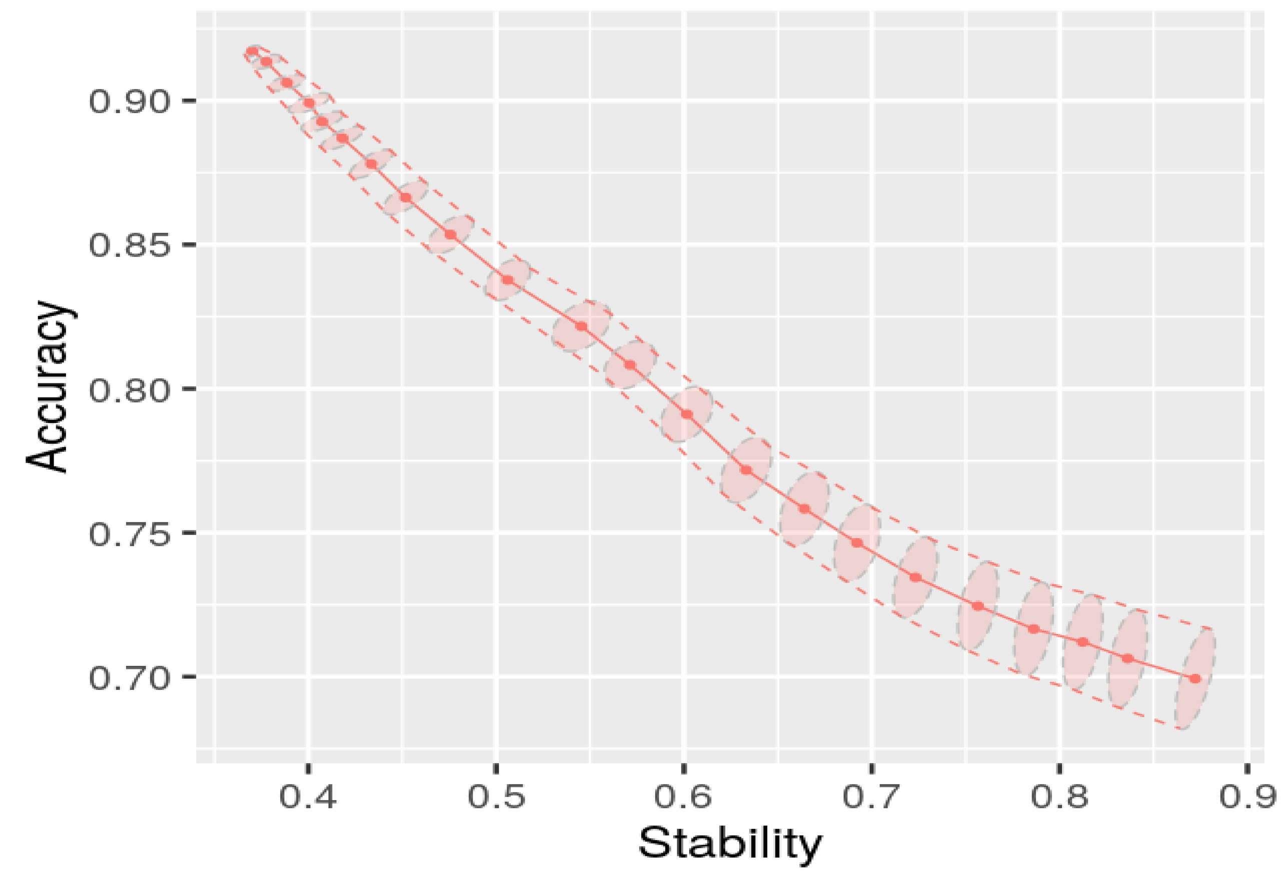
- Drops a fraction of the least significant features at each step.
- Until the desired number of features (k) is met.
- A feature f with a lower β_f has a higher probability to be selected and vice-versa ⇒ control the accuracy-stability tradeoff by tuning β.
- Paper: $\beta_f \sim \Gamma(\alpha, 1)$
- Results on Prostate (n=102, d=12600, k=20):



- **Domain experts** can thus tune α and choose any Pareto-optimal compromise.

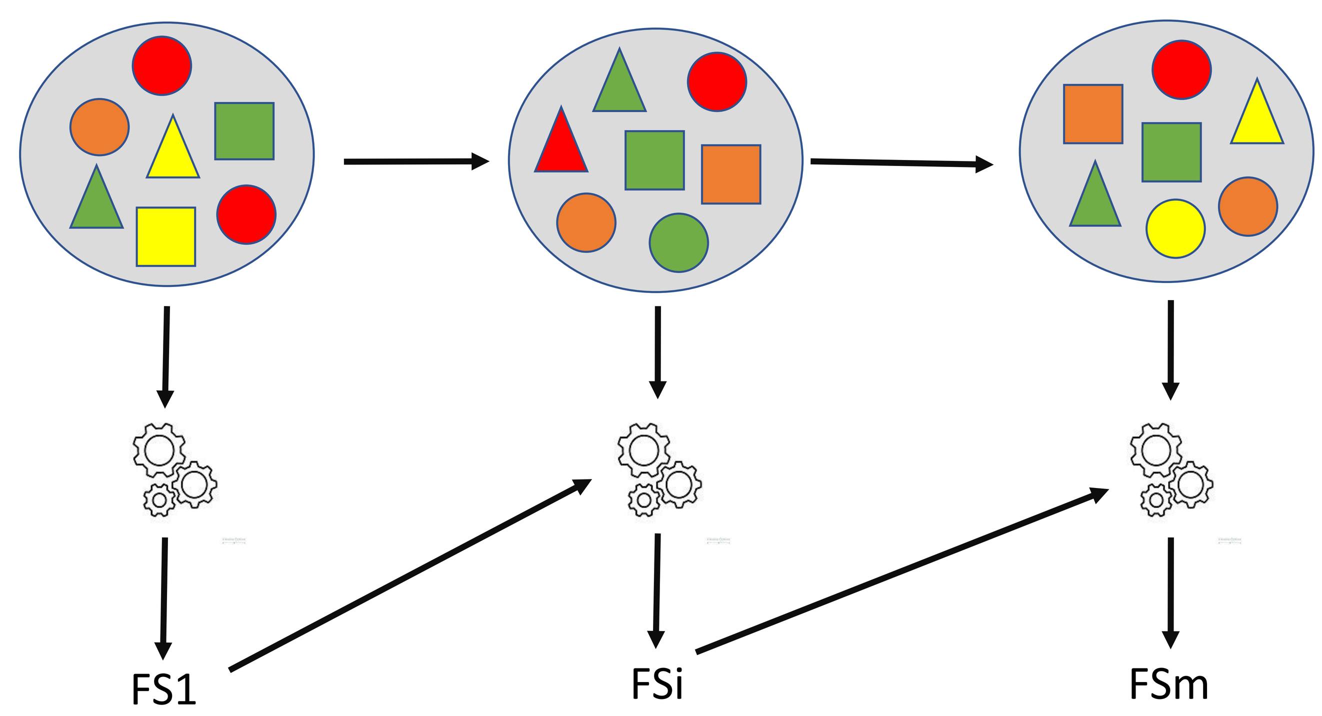
Confidence intervals on HV

- Possible (see paper) to define ellipsoidal confidence regions for each Pareto-optimal trade-off ⇒ use the most dominated and most dominant point of each ellipse to compute the bounds of the CI.

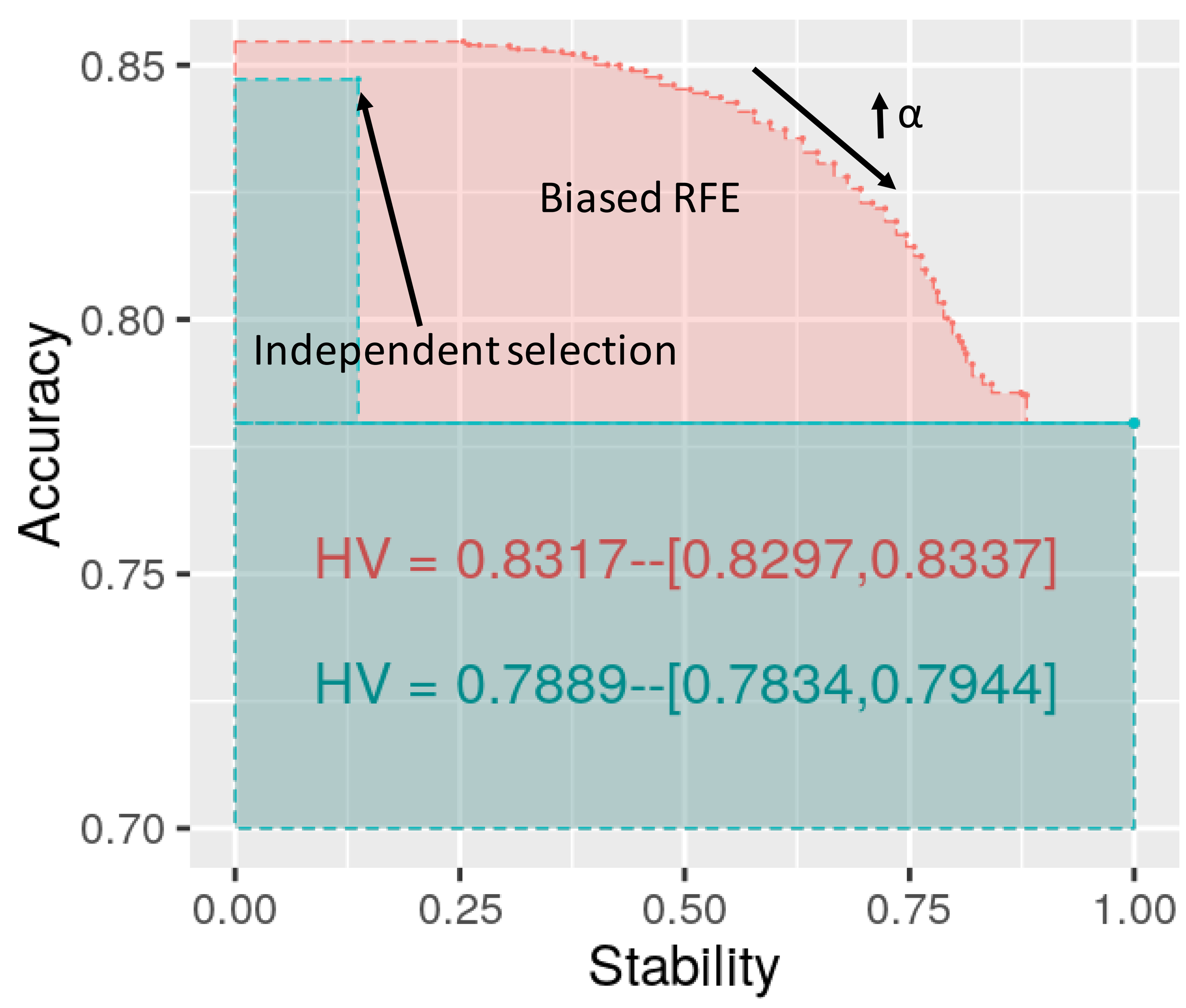


Transfer learning

- Sometimes, one wants to find similar feature subsets for different tasks.
- **Transfer learning**: tasks are ordered



- Stability increase if feature f is taken at task number i: $2p_f - 1$ with p_f the selection frequency of feature f in task $[0, i[$.
- Paper: $\beta_f \propto \exp(-\alpha * p_f)$ ⇒ prioritize more features which selection would increase more the stability.



Future work

- Extension to multi-task selection.
- Apply differential shrinkage to other losses or regularizations (Elastic Net penalty, deep feature selectors, ...).

References

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown.
On the stability of feature selection algorithms.
The Journal of Machine Learning Research, 18(1):6345–6398, 2017.

victor.hamer@uclouvain.be