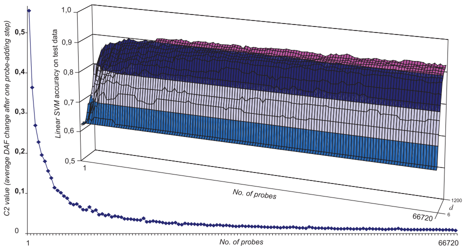
On Stopping Rules in Dependency-Aware Feature Ranking

Petr Somol[1,2] **Jiří Grim**[1] **Jiří Filip**[1] **Pavel Pudil**[2]

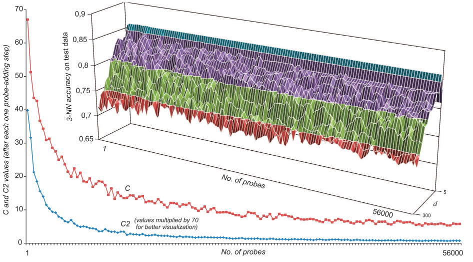
[1] Institute of Information Theory and Automation of the AS CR
[2] Faculty of Management, Prague University of Economics, Czech Republic
{somol,grim,filip,pudil}@utia.cas.cz

Feature Selection in very-high-dimensional or small sample problems is particularly prone to computational and robustness complications. It is common to resort to feature ranking approaches only or to randomization techniques. A recent novel approach to the randomization idea in form of Dependency-Aware Feature Ranking (DAF) has shown great potential in tackling these problems well. Its original definition, however, leaves several technical questions open. In this paper we address one of these questions: how to define stopping rules of the randomized computation that stands at the core of the DAF method. We define stopping rules that are easier to interpret and show that the number of randomly generated probes does not need to be extensive.

Reuters data - SVM Classifier accuracy and $C2$ convergence during DAF probe generation



Madelon data - 3-NN Classifier accuracy and C and $C2$ convergence during DAF probe generation



Motivation

Feature selection (FS):

- one of dimensionality reduction techniques
- preserves meaning of the selected original data features
- irrelevant features are discarded
- **Problem formulation**: N-dimensional feature space → classification of objects described by means of features $f_1, f_2, \dots f_N$ → finite number of mutually exclusive classes
- **Task**: optimal selection of subset of the most informative features
- **Solution**: choice of criterion and search algorithm → *When to stop a time demanding search?*

Context to Related Work

Recent shift from low-to-mid dimensional problems to very-high-dimensional ones.
Related FS problems:

- insufficient sample size
- computational complexity

Solutions:

- *feature ranking*: ignoring inter-feature dependencies → overfitting
- *randomized methods* (Relief algorithm, Genetic algorithm, Simulated annealing, Tabu search) → user-defined termination of search process.
- combination of both → *Dependency-aware feature ranking (DAF)* [1] → evaluate features contributions in a sequence of randomly generated feature subsets → *when to stop the random subsets generation?*

Dependency-Aware Feature Ranking

Assume set of all features $F = \{f_1, f_2, \dots, f_N\}$ for each subset of features $S \subset F$ a feature selection criterion $J(\cdot)$ is evaluated to measure the quality of S

Generate sequence of random subsets - *probes* $\mathbb{S} = \{S_1, S_2, \dots, S_K\}$, $S_j \subset F, \quad j = 1, 2, \dots, K$, [1]
Sufficiently large sequence of feature subsets $\mathbb{S}$ → utilize the information contained in the criterion values $J(S_1), J(S_2), \dots, J(S_K)$ to assess how each feature adds to the criterion value. We compare the quality of probe subsets containing f vs. quality of probe subsets not including f.
Mean quality μ_f of subsets $S \in \mathbb{S}$ containing the considered feature

$$\mu_f = \frac{1}{|\mathbb{S}_f|} \sum_{S \in \mathbb{S}_f} J(S), \quad \mathbb{S}_f = \{S \in \mathbb{S} : f \in S\} \tag{1}$$

mean quality $\bar{\mu}_f$ of subsets $S \in \mathbb{S}$ not containing the considered feature f:

$$\bar{\mu}_f = \frac{1}{|\bar{\mathbb{S}}_f|} \sum_{S \in \bar{\mathbb{S}}_f} J(S), \quad \bar{\mathbb{S}}_f = \{S \in \mathbb{S} : f \notin S\} \tag{2}$$

Quality of the feature f is expressed by coefficient:

$$DAF(f) = \mu_f - \bar{\mu}_f, \quad f \in F. \tag{3}$$

Design of Novel Stopping Rules

Stopping condition 1: *Change of Feature Order.*
When adding probes to $\mathbb{S}$ keep updating the feature ordering according to DAF(f) → evaluate changes in feature's order. Defining a threshold on the change → allow to stop adding probes when the ordering is not changing substantially any more.
$C[\mathbb{S}_1, \mathbb{S}_2] = \frac{1}{N} \sum_{f=1}^{N} |DAF(f)_{idx}^{\mathbb{S}_1} - DAF(f)_{idx}^{\mathbb{S}_2}|$

Stopping condition 2: *Change of Average DAF value.*
When adding probes to $\mathbb{S}$ keep updating the feature ordering according to DAF(f) → evaluate changes in feature's order. These changes decrease with increasing number of probes → define a threshold on DAF value change to specify when the change is to be considered small enough to justify stopping the process.$C2[\mathbb{S}_1, \mathbb{S}_2] = \frac{1}{N} \sum_{f=1}^{N} |DAF(f)^{\mathbb{S}_1} - DAF(f)^{\mathbb{S}_2}|$

Stopping condition 2a:
Relative Change of Average DAF value.
Probes adding and recalculating DAF coefficients for each feature after the additions leads to changes in DAF coefficient value for some or all features. Stop probe adding when for the k-th added probe it is true that

$$\frac{C2[\mathbb{S}_k, \mathbb{S}_{k+1}]}{C2[\mathbb{S}_1, \mathbb{S}_2]} < t$$

for a pre-specified threshold t.

Experimental Evaluation

Reuters-21578 text categorization benchmark

- 33 classes, 10105 features
- http://www.daviddlewis.com/resources/testcollections/reuters21578

Madelon artificial data [2]

- 2 classes, 500 features (20 informative, 480 noise)

Experiment setup [1]:

- accuracy on Reuters data evaluated by SVM
- accuracy on Madelon data evaluated by 3-NN
- C and $C2$ → computed after each 400-th probe
- probe size was limited to 200 features

Results (figures above):

- quick improvement of classification accuracy after a small number of initially added probes
- subset sizes (Madelon: $d = 20$, Reuters: $d = 1000$) represent the most informative features

Conclusions

- Proposed alternative stopping rules in Dependency-Aware Feature Ranking.
- Thresholding the averaged change in DAF value (when adding random probe subsets to the evaluated sequence) is preferable to other stopping rules in terms of interpretability, especially for unknown underlying data.
- DAF is fairly robust and does not require excessive numbers of randomized probes.

References

[1] Somol P., Grim J., Pudil P.: *Fast dependency-aware feature selection in very-high-dimensional pattern recognition.* In: Proceedings of the IEEE Inter. Conf. on Systems, Man and Cybernetics (SMC), 2011, pp. 502-509

[2] Newman D., Hettich S., Blake C., Merz C.: *UCI repository of machine learning databases* (1998)

CIARP 2013, Havana, Cuba