Feature Subset Selection Using the Wrapper Approach: Dynamic Search Space Topology

Ron Kohavi and Dan Sommerfield
Stanford University

KDD-95

(ronnyk@CS.Stanford.EDU)

Motivation

Feature subset selection (FSS) is the process of selecting a subset of features to show the induction algorithm. Reasons for doing FSS:

1. Improve accuracy. Many induction algorithms degrade in performance when given too many features.

2. Improve comprehensibility.

3. Reduce measurement cost: measuring features may cost money.

Talk Outline

☞ ① Feature subset selection and the wrapper approach.

② Compound operators.

③ Experimental results.

④ Summary.

Optimal Features

Given an induction algorithm, $\mathcal{I}$, and a dataset, D, the *optimal feature subset*, S^*, is the set of features such that the generated classifier has the highest prediction accuracy.

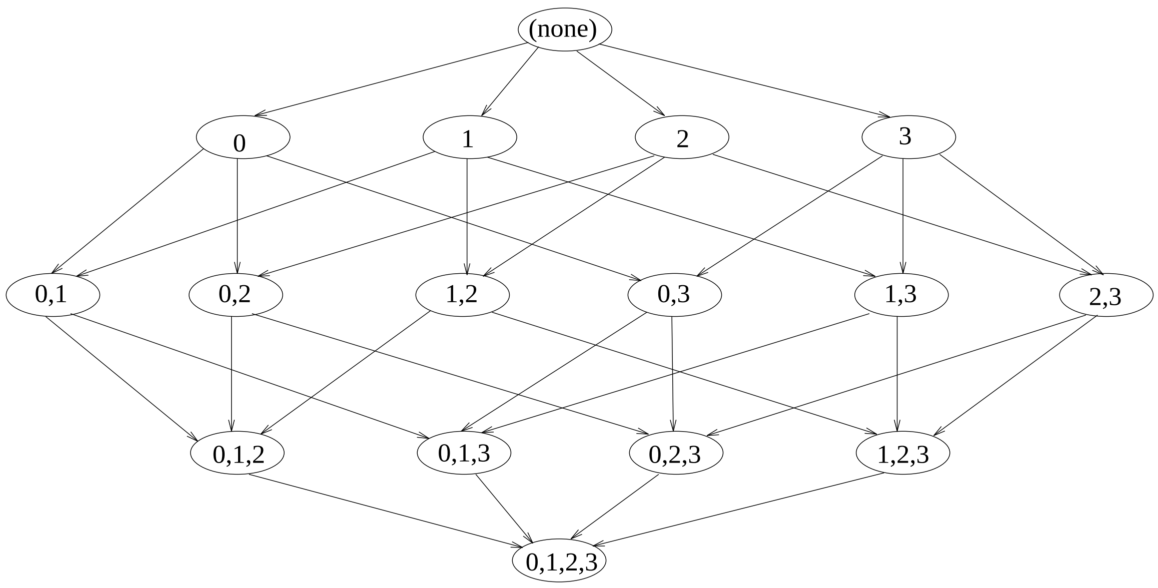
$$S^* = \operatorname*{arg\,max}_{S' \subseteq S} \ \mathrm{acc}(\mathcal{I}(D_{S'}))$$

where $\mathcal{I}(D_{S'})$ is the classifier built by $\mathcal{I}$ from the dataset D using only features in S'.

FSS as State Space Search

FSS can be described as state space search.

1. Each node (state) represents a feature subset.
2. The value of a node is the estimated prediction accuracy.
3. The operators are commonly add/delete feature.

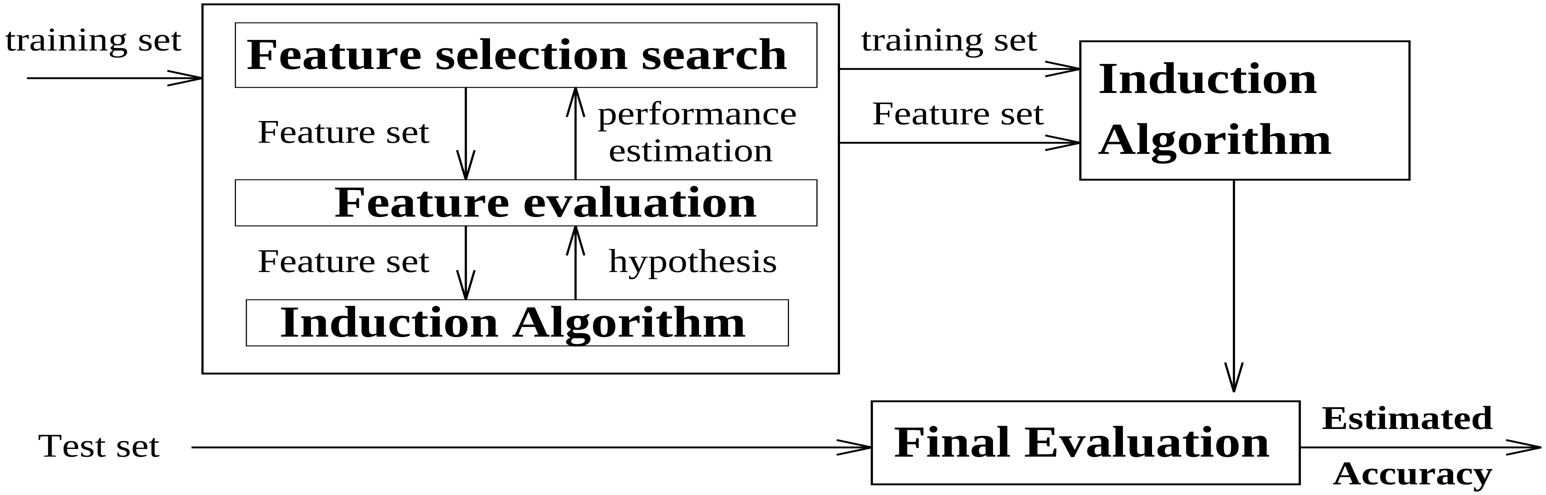


The Wrapper Approach

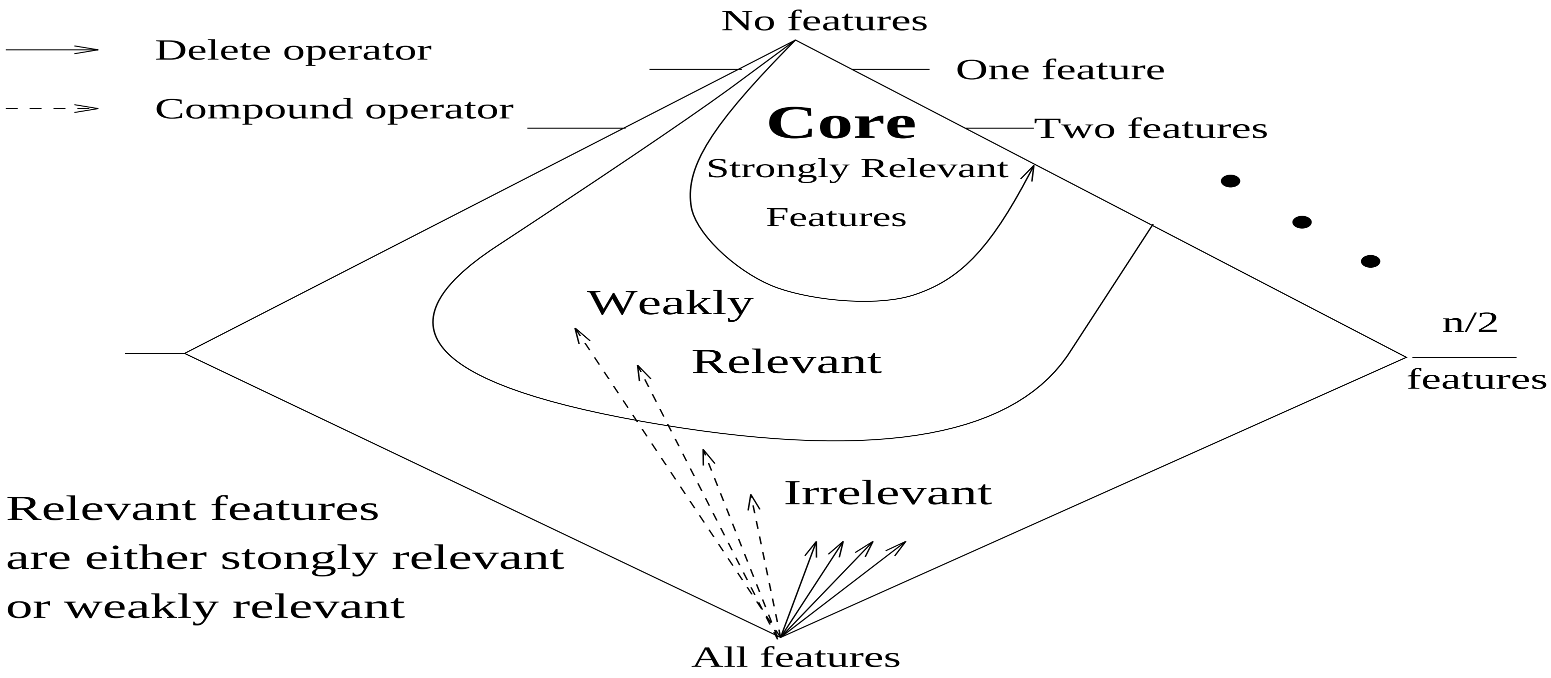
1. In the *wrapper approach*, we use the induction algorithm as a black box (whereas filter approaches use just the data).

2. A search is conducted in the space of subsets with add/delete operators (we used best-first search).

3. The heuristic for the search is the estimated prediction accuracy using cross-validation.



Compound Operators : Motivation



The state space. If a feature subset contains an irrelevant feature, it is in the irrelevant area; if it contains only core features it is in the core region; otherwise, it is in the relevant region. The dotted arrows indicate compound operators.

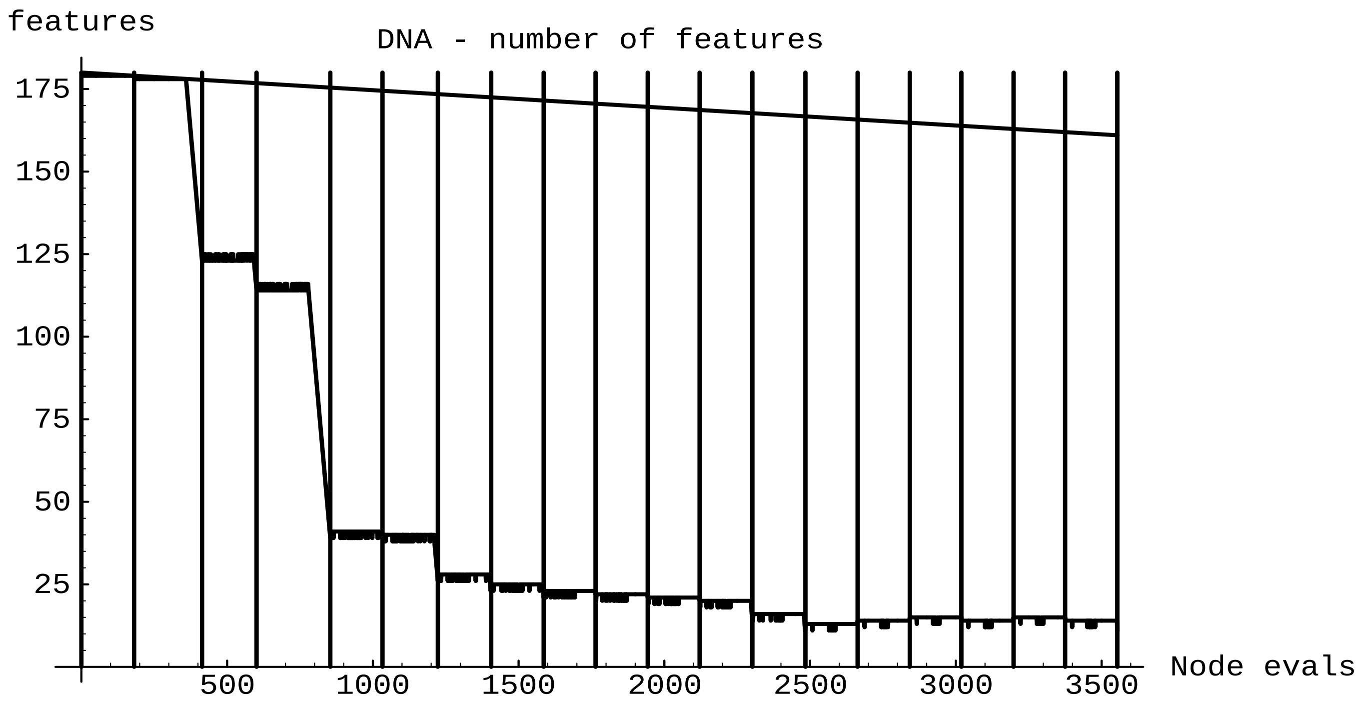
Compound Operators

Compound operators are operators that are dynamically created *after* the standard set of children, created by the add and delete operators, have been evaluated.

Compound operators combine operators that led to the best children into a single dynamic operator.

If we rank the operators by the estimated accuracy of the children, then we can define compound operator c_i to be the combination of the best $i + 1$ operators. For example, the first compound operator will combine the best two operators.

DNA Run on C4.5



DNA: Number of features evaluated as the search progresses (C4.5, BFS, compound backward). The vertical lines signify a node expansion, where the children of the best node are expanded.

(ronnyk@CS.Stanford.EDU)

Experimental results

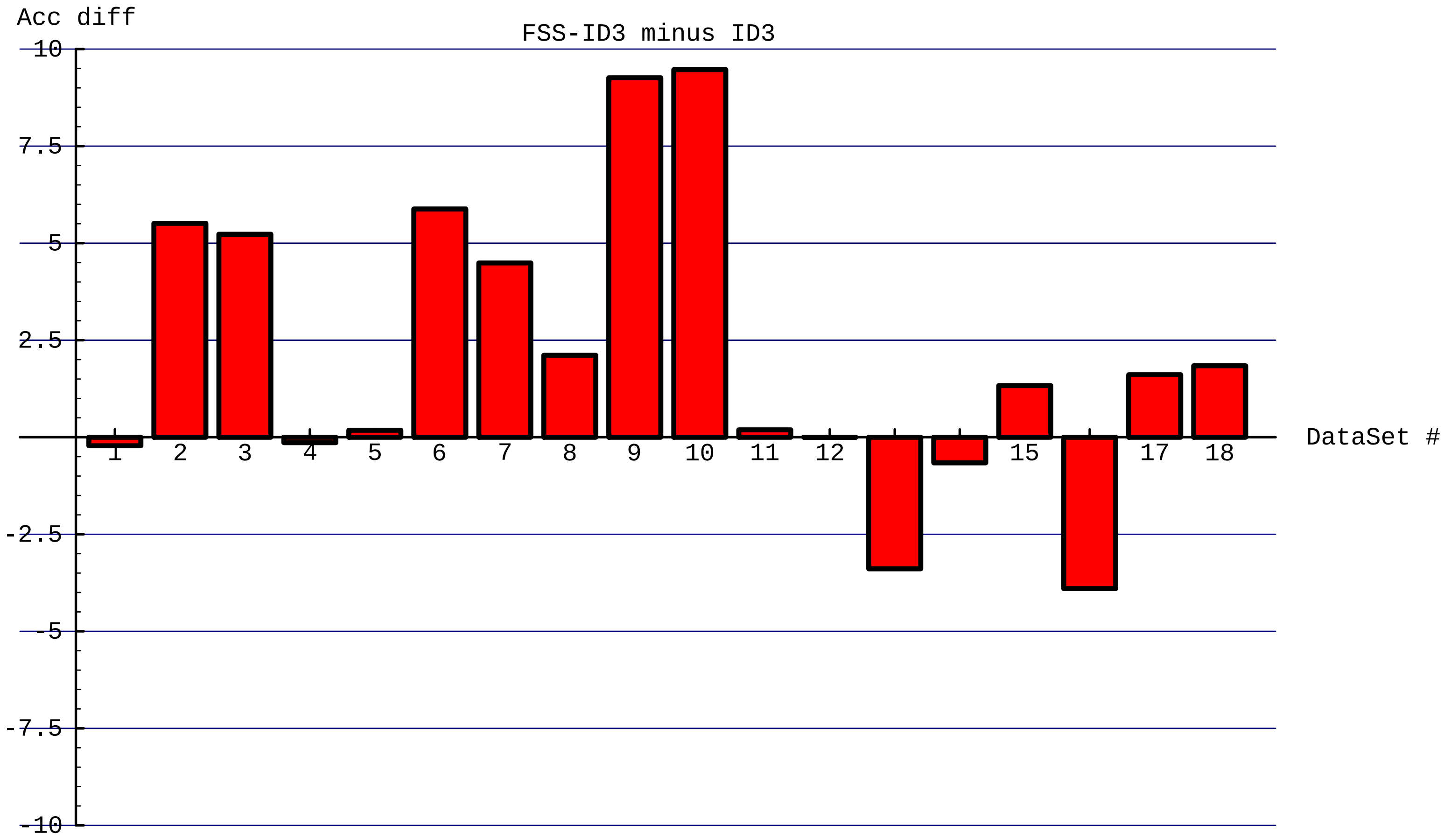
We ran the wrapper over ID3 and Naive-Bayes. The runs represent best-first search starting with the empty set of features (forward selection) and compound operators.

ID3 is a top-down induction of decision trees, but with no pruning. The FSS not only removes bad features to split on, but also provides a pruning mechanism.

Naive-Bayes computes the probability of each class given the instances, assuming conditional independence of the features given the class.

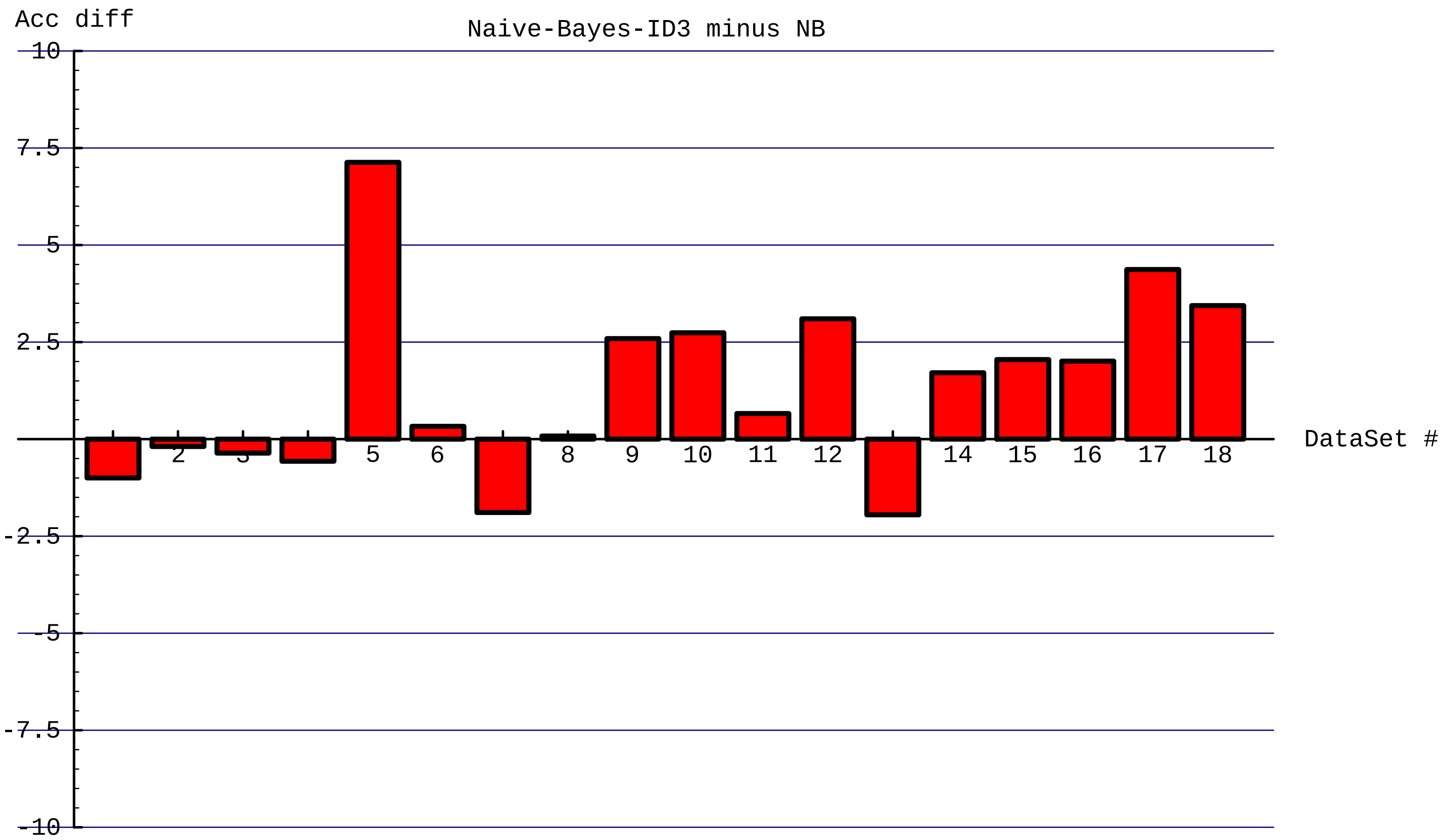
Final feature subsets are evaluated on unseen test instances using 5-fold cross-validation.

ID3 with FSS versus ID3

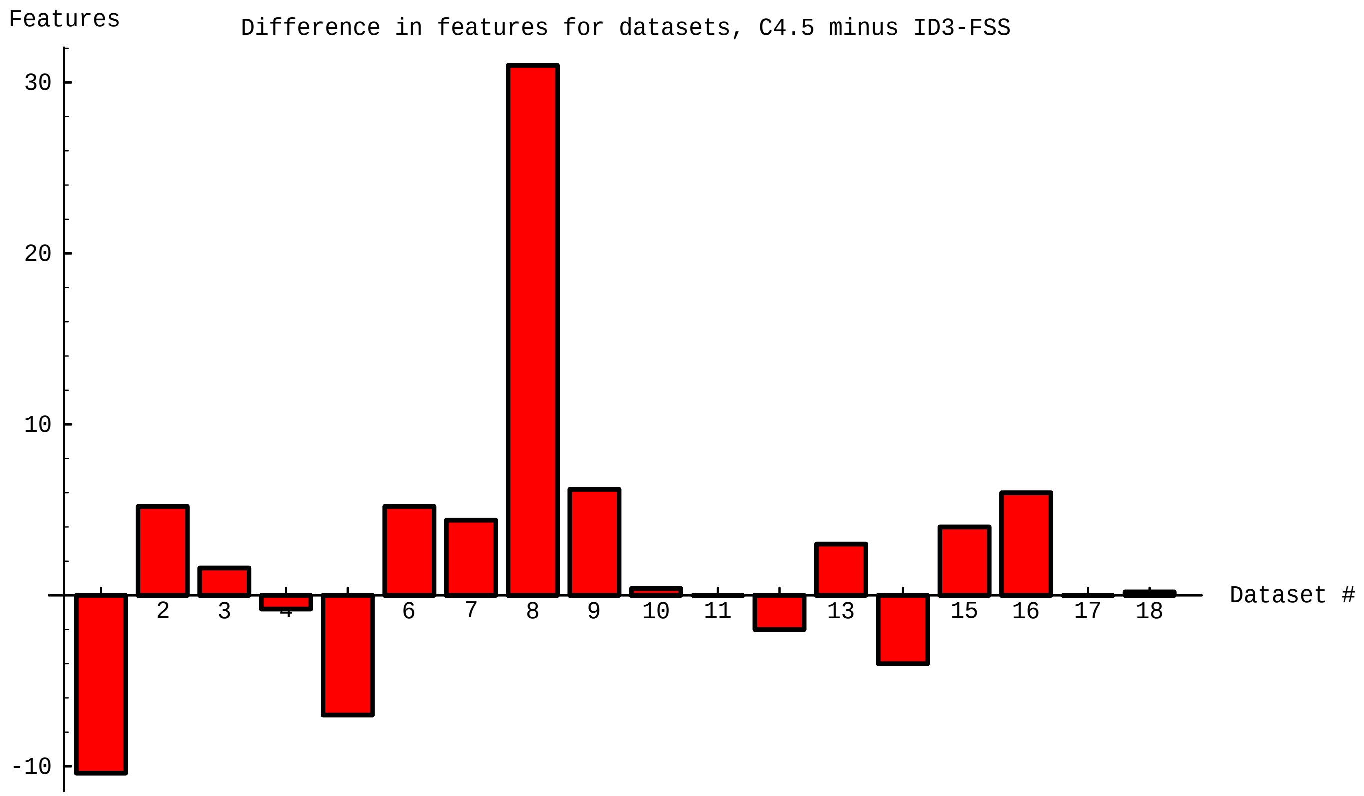


(ronnyk@CS.Stanford.EDU)

Naive-Bayes with FSS versus NB



The Number of Features



The difference in number of features used by C4.5 and ID3-FSS. Positive means C4.5 is using more features.

Summary

1. The wrapper approach was reviewed. The idea is to wrap around an existing learning algorithm, rather than use some statistical measure that may be inappropriate.

2. Compound operators reduce the number of node evaluations.

3. Backward search is now feasible, and results are slightly better. For example, Naive-Bayes on the StatLog DNA achieves 96.1% accuracy, higher than the 23 algorithms tested in StatLog.

4. Problems: (i) very slow; (ii) overfitting (in paper).