Irrelevant Features and the Subset Selection Problem

George John, Ronny Kohavi, and Karl Pfleger

Computer Science Department,
Stanford University

ML-94

Framework & Talk Outline

Framework: Supervised classification learning.

Input:	A training set consisting of labelled instances. Each instance is a list of feature values.
Output:	A classifier that should perform well on future data.

Outline:

1. Problem definition.
2. What is relevance and irrelevance.
3. Experiments using the "wrapper model."

Feature Subset Selection

Problem: Select a subset S^* from the set of features S, such that a given performance measure $f(S^*)$ is minimal. Formally:

$$S^* = \arg\min_{S' \subseteq 2^S} f(S')$$

Examples of performance functions are:

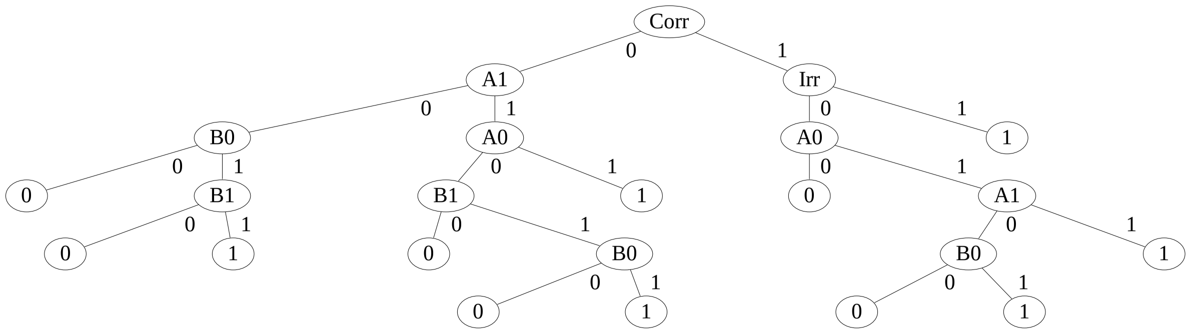
1. Error rate, *i.e.*, error of induced classifier on unseen test set.
2. Error rate plus a function of the size of the induced structure.
3. Number of features (under some restrictions on subsets).

Motivating Example : Corral

There are six Boolean variables: $A_0, A_1, B_0, B_1,$ *Irr*, *Corr*. *Irr* is uniformly random, and *Corr* correlates with the correct label 75% of the time. The concept is

$$(A_0 \wedge A_1) \vee (B_0 \wedge B_1)$$

ID3 induces the following tree on a training set of size 32



Relevance & Irrelevance

Intuitively, we want to remove "irrelevant" features. But what are irrelevant features?

Almuallim & Dietterich 1990 A feature X_i is said to be **relevant** to a concept $\mathcal{C}$ if X_i appears in every Boolean formula that represents $\mathcal{C}$ and irrelevant otherwise.

Probabilistic X_i is relevant if there exists an assignment to the variables such that

$$p(Y = y \mid x_1, \ldots, x_n) \neq p(Y = y \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

Example: In a domain with three Boolean variables, there exists a domain constraint $X_3 = \overline{X}_2$, and the concept is $Y = X_1 \oplus X_2$.

Strong & Weak Relevance

We will adopt the probabilistic definition previously described, as defining **strong relevance.**

Weak relevance: A feature X_i is weakly relevant iff it is not strongly relevant, and there exists a subset of features S_i' such that knowing the value of X_i changes the conditional probability:

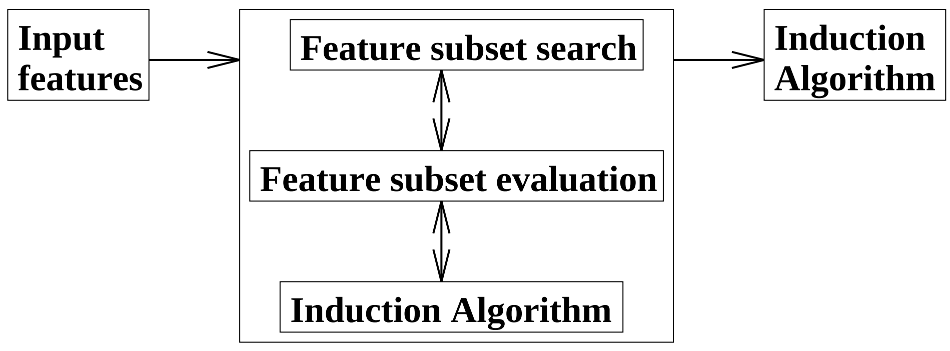
$$p(Y = y \mid X_i = x_i, S_i' = s_i') \neq p(Y = y \mid S_i' = s_i')$$

Filter Model

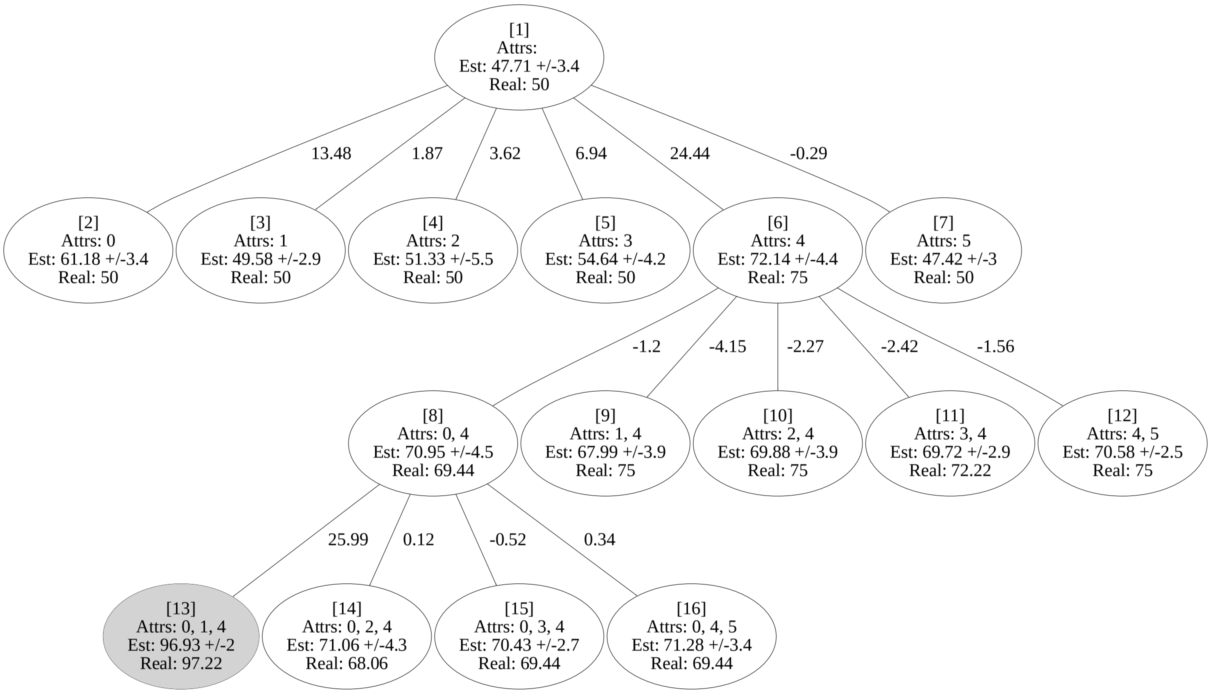
1. The common model for feature subset selection is the *filter model*.
2. In this model, the feature subset selection is done as a pre-processing step.
3. Problem: ignores induction algorithm.
4. Examples: Statistical measures, Focus, Relief, Cardie.

Wrapper Model

1. In the *wrapper model*, we use the induction algorithm as a black box.
2. A search is conducted in the space of subsets.
3. The accuracy is estimated using cross-validation or Bootstrap.



Example of Forward Selection



Experimental Results

1. Search method: hill-climbing, add-feature and delete-feature as operators.
2. Initial states: no features and all features.
3. Evaluation: 25-fold Cross Validation.
4. Observations:
   (a) Improved accuracy on Corral from 82% to 100%, and parity from 50% to 100%.
   (b) No significant difference on most datasets, but trees were smaller. For example, *Credit* was 16 nodes vs. 44.
   (c) RelieveD (deterministic Relief) removes features, but not enough.

Discussion

1. Cross validation has very high variance in its estimates. This seems to be the biggest problem.
2. The search method was very limited (hill-climbing). Better search methods such as Best-First Search can be used.
3. Forward selection seems to work better for “real” problems. It is faster and finds smaller structures.
4. Subsets found by RelieveD are good, but sometimes too large. This may serve as a good starting point for backward elimination.

Summary

1. Finding a good feature subset is an important problem for real datasets. A “good” subset can improve performance and comprehensibility.
2. We defined strong & weak relevance, and showed that one should be looking for all strongly relevant features, non of the irrelevant features, and a “good” subset of the weakly relevant features.
3. Experimental results show improved performance in some cases, but usually no significant change. The induced trees are smaller and thus easier to understand.