Overview

- *Feature Selection* is a process that chooses an optimal subset of features *from original set of features*

Overview

- Reasons for performing FS may include:
  - removing irrelevant data.
  - increasing predictive accuracy of learned models.
  - improving learning efficiency, such as reducing storage requirements and computational cost.
  - reducing the complexity of the resulting model description, improving the understanding of the data and the model.

Perspectives

1. searching for the best subset of features.

2. criteria for evaluating different subsets.

Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination

Random Search

Non-deterministic or random: No predefined way to select feature candidate (i.e., probabilistic approach)

- Optimal subset depends on the number of trials
- Need more user-defined parameters

Example GA for feature selection

Perspectives: Selection Criteria

– Information Measures.

- Shannon's Entropy:

$$-\sum_{i} P(c_i) \log_2 P(c_i).$$

- Information gain:

$$IG(A) = I(D) - \sum_{j=1}^{p} \frac{|D_j|}{|D|} I(D_j^A)$$

Perspectives: Selection Criteria

- Distance Measures.
  - Measures of separability, discrimination or divergence measures . Minkowski distance is a generalized distance metric as given below.

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

  - Some common values of ‘p’ are:-
  - p = 1, Manhattan Distance
  - p = 2, Euclidean Distance

Perspectives: Selection Criteria

- Dependence Measures.
  - known as measures of association or correlation.
  - Its main goal is to quantify how strongly two variables are correlated or present some association with each other, in such way that knowing the value of one of them, we can derive the value for the other.
  - *Pearson correlation* coefficient:

$$\rho(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

Perspectives: Selection Criteria

- Consistency Measures.
  - They attempt to find a minimum number of features that separate classes.
  - An inconsistency is defined as the case of two examples with the same inputs (same feature values) but with different output feature values (classes in classification).

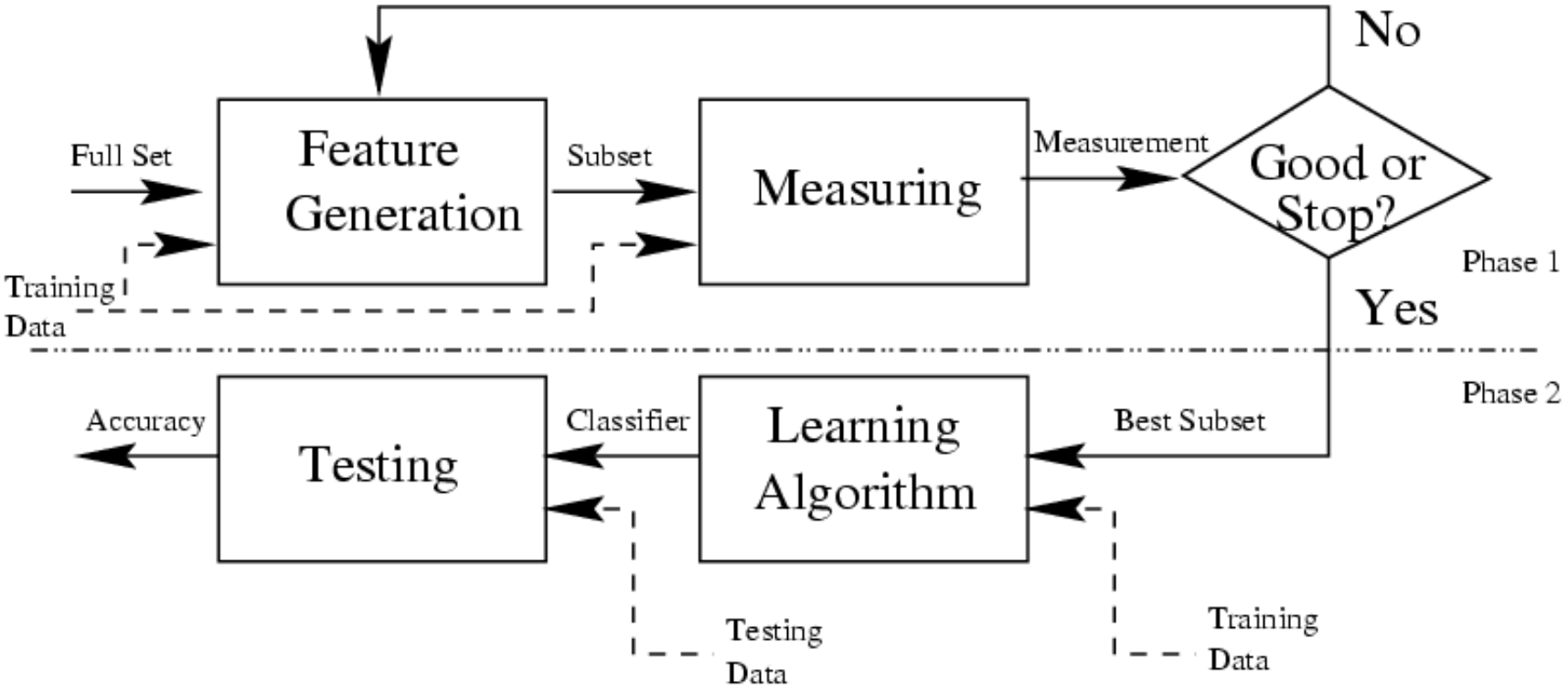
Perspectives: Selection Criteria

- Accuracy Measures.
  - This form of evaluation relies on the classifier or learner. Among various possible subsets of features, the subset which yields the best predictive accuracy is chosen

Models of Feature Selection

- Filter model
  - Separating feature selection from classifier learning
  - Relying on general characteristics of data (*information, distance*, *dependence*, *consistency*)
  - No bias toward any learning algorithm, fast
- Wrapper model
  - Relying on a predetermined classification algorithm
  - Using predictive accuracy as goodness measure
  - High accuracy, computationally expensive

Filter Model



Wrapper Model