# Downsizing Data for High Performance in Learning

## - Introduction to Feature Selection Methods

**Huan Liu and Robert Stine**

**Arizona State Univ and Univ of Penn**


# Outline

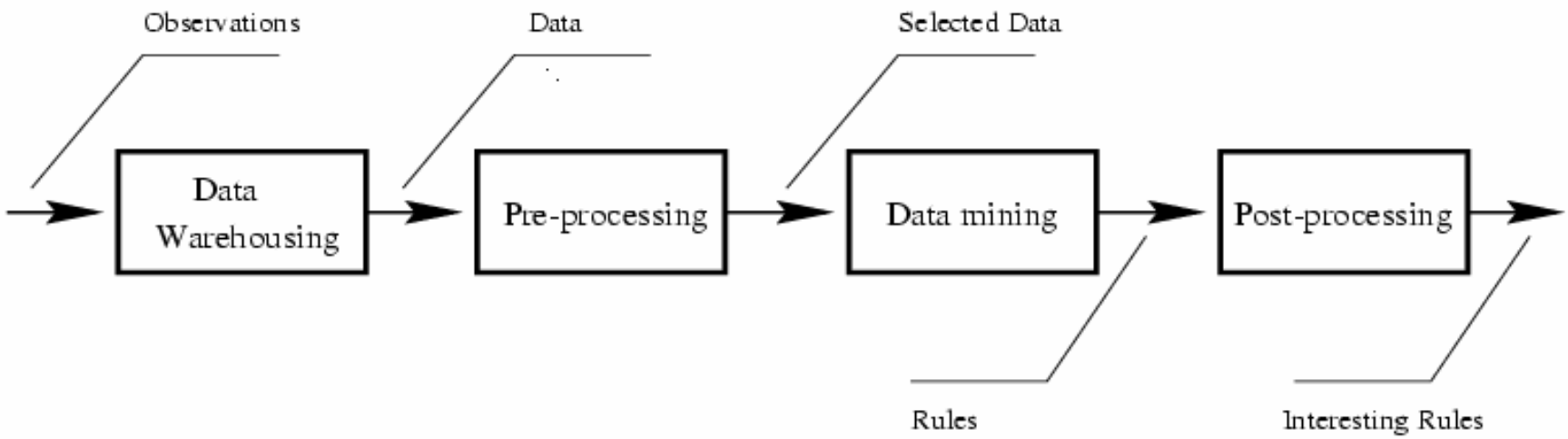
- Introduction and Basics (I)
- Evaluation and Model-based Methods (II)
- `Wide' Data and Feature Redundancy (III)
- Adaptive Selection and Sequential Testing (IV)

# Why Feature Selection?

- It is so easy and convenient to collect data
  - An experiment
- Data is not collected only for data mining
- Data accumulates in an unprecedented speed
- Data preprocessing is an important part for *effective* machine learning and data mining
- Feature selection is an effective approach to downsizing data

# A General Model of KDD

- KDD process




- Data mining
  - Applying analytical methods and tools to data to identify patterns, statistical or predictive models, and relationships among massive data

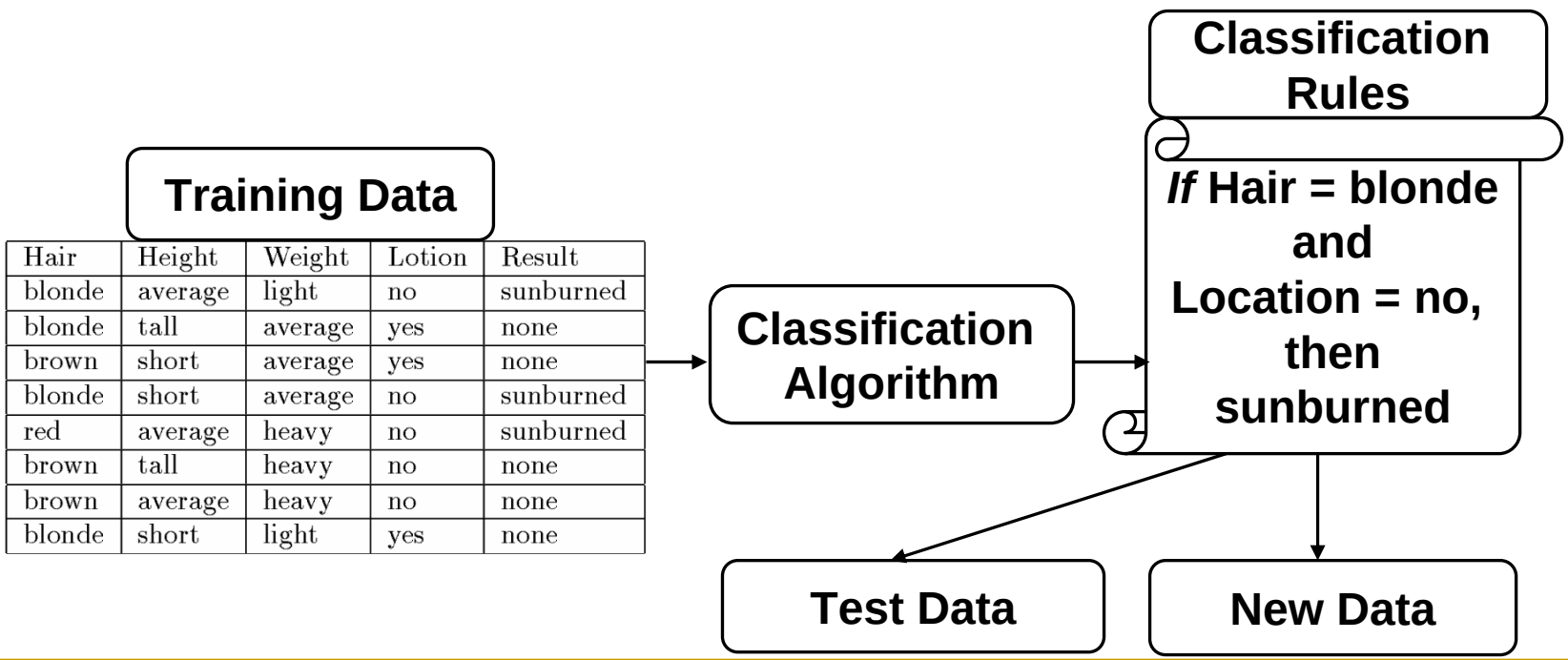
# Data Format

- Features/attributes
  - Discrete (nominal, ordinal)
  - Continuous
- Instances, tuples, examples, or data points
- An example of feature-based data (sunburn)

| Hair | Height | Weight | Lotion | Result |
|---|---|---|---|---|
| blonde | average | light | no | sunburned |
| blonde | tall | average | yes | none |
| brown | short | average | yes | none |
| blonde | short | average | no | sunburned |
| red | average | heavy | no | sunburned |
| brown | tall | heavy | no | none |
| brown | average | heavy | no | none |
| blonde | short | light | yes | none |

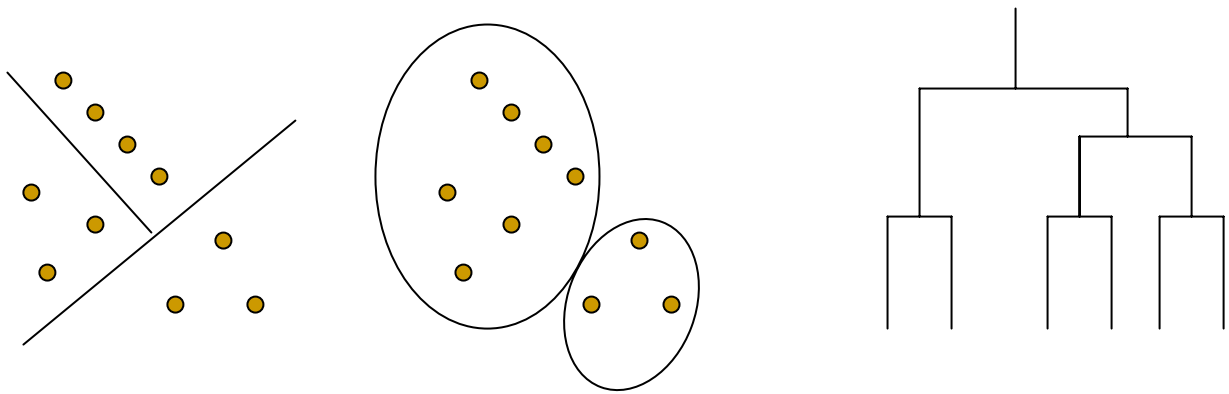
# Classification

- A process of predicting the classes of unseen instances based on patterns learned from available instances
- Supervised learning with labeled data



# Clustering

- A process of grouping objects (or instances) into *clusters* so that objects are similar to one another within a cluster but dissimilar to objects in other clusters
- Unsupervised learning with unlabeled data
- Clustering tasks



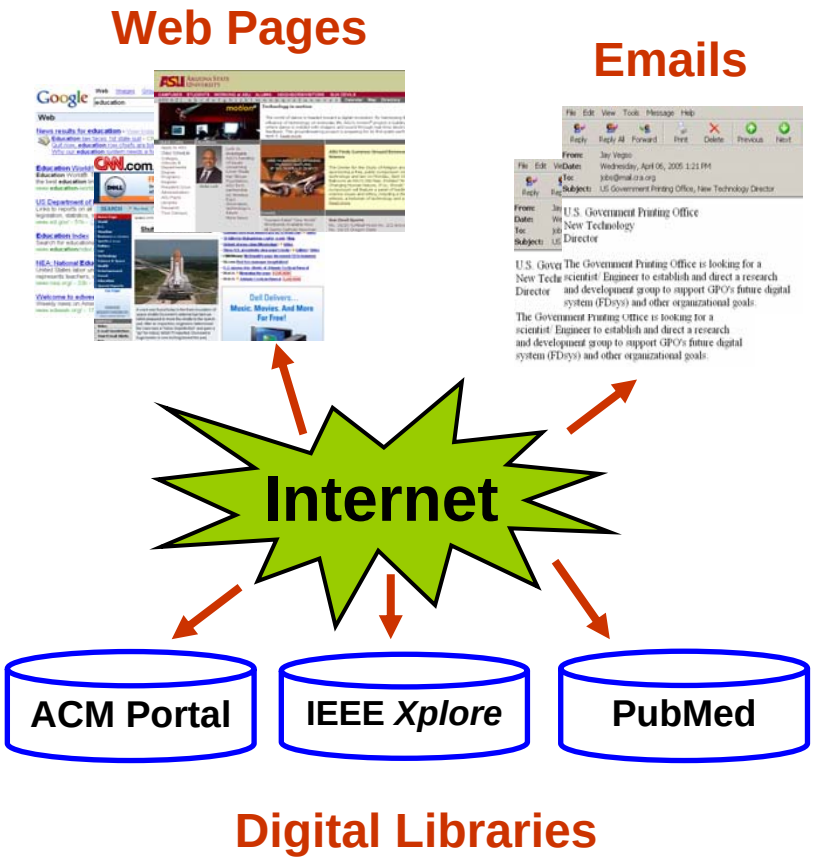
# Feature Selection

- Feature selection
  - A process that chooses an optimal subset of features according to a certain criterion
- Objectives
  - To reduce dimensionality and remove noise
  - To improve learning performance
    - Speed of learning
    - Predictive accuracy
    - Simplicity and comprehensibility of learned results

# Examples of Feature Selection

- Predicting Credit Risk
- Customer Relationship Management
- Text categorization
- Microarrray data analysis

# Online Document Classification

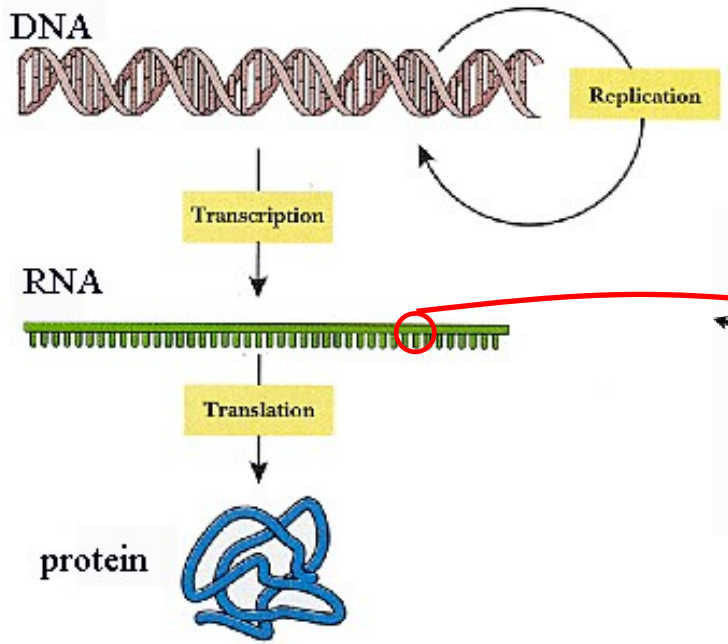



| | Terms | | | | |
|---|---|---|---|---|---|
| | $T_1$ | $T_2$ | ..... | $T_N$ | C |
| $D_1$ | 12 | 0 | ..... | 6 | Sports |
| $D_2$ | 3 | 10 | ..... | 28 | Travel |
| ⋮ | ⋮ | | | ⋮ | ⋮ |
| $D_M$ | 0 | 11 | ..... | 16 | Jobs |

Documents

- **Task:** To classify unlabeled documents into categories
- **Challenge:** thousands of terms
- **Solution:** to apply feature selection

# Gene Expression Microarray Analysis




Expression Microarray

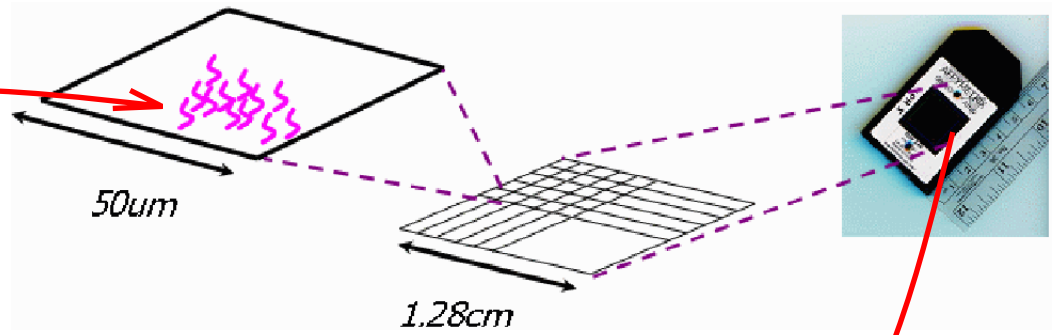



Image Courtesy of Affymetrix

- **Task:** To classify novel samples into known disease types (disease diagnosis)
- **Challenge:** thousands of genes, few samples
- **Solution:** to apply feature selection

| Sample \ Gene | M23197_at | U66497_at | M92287_at | . . . | Class |
|---|---|---|---|---|---|
| Sample 1 | 261 | 88 | 4778 | . . . | ALL |
| Sample 2 | 101 | 74 | 2700 | . . . | ALL |
| Sample 3 | 1450 | 34 | 498 | . . . | AML |
| . | . | . | . | . . . | . |
| . | . | . | . | . . . | . |
| . | . | . | . | . . . | . |

Expression Microarray Data Set

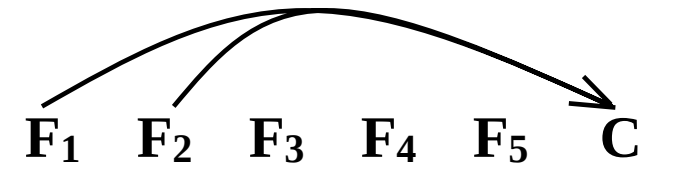
# Basics and Algorithms

- Definitions of subset optimality
- Perspectives of feature selection
  - Subset search and feature ranking
  - Feature/subset evaluation measures
  - Feature relevance and redundancy
  - Models: filter vs. wrapper
  - Results validation and evaluation
- Representative algorithms for classification
- Selection of algorithms

# Subset Optimality for Classification

- A *minimum* subset that is sufficient to construct a hypothesis consistent with the training examples (*Almuallim & Dietterich 1994*)
  - Optimality is based on the training set
  - The optimal set may *overfit* the training data
- A *minimum* subset $G$ such that $\mathbf{P}(C|G)$ is equal or as close as possible to $\mathbf{P}(C|F)$ (*Koller & Sahami 1996*)
  - Optimality is based on the entire population
  - But, only the training part of the data is available

# An Example for Optimal Subset



| $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ | C |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 | 1 |

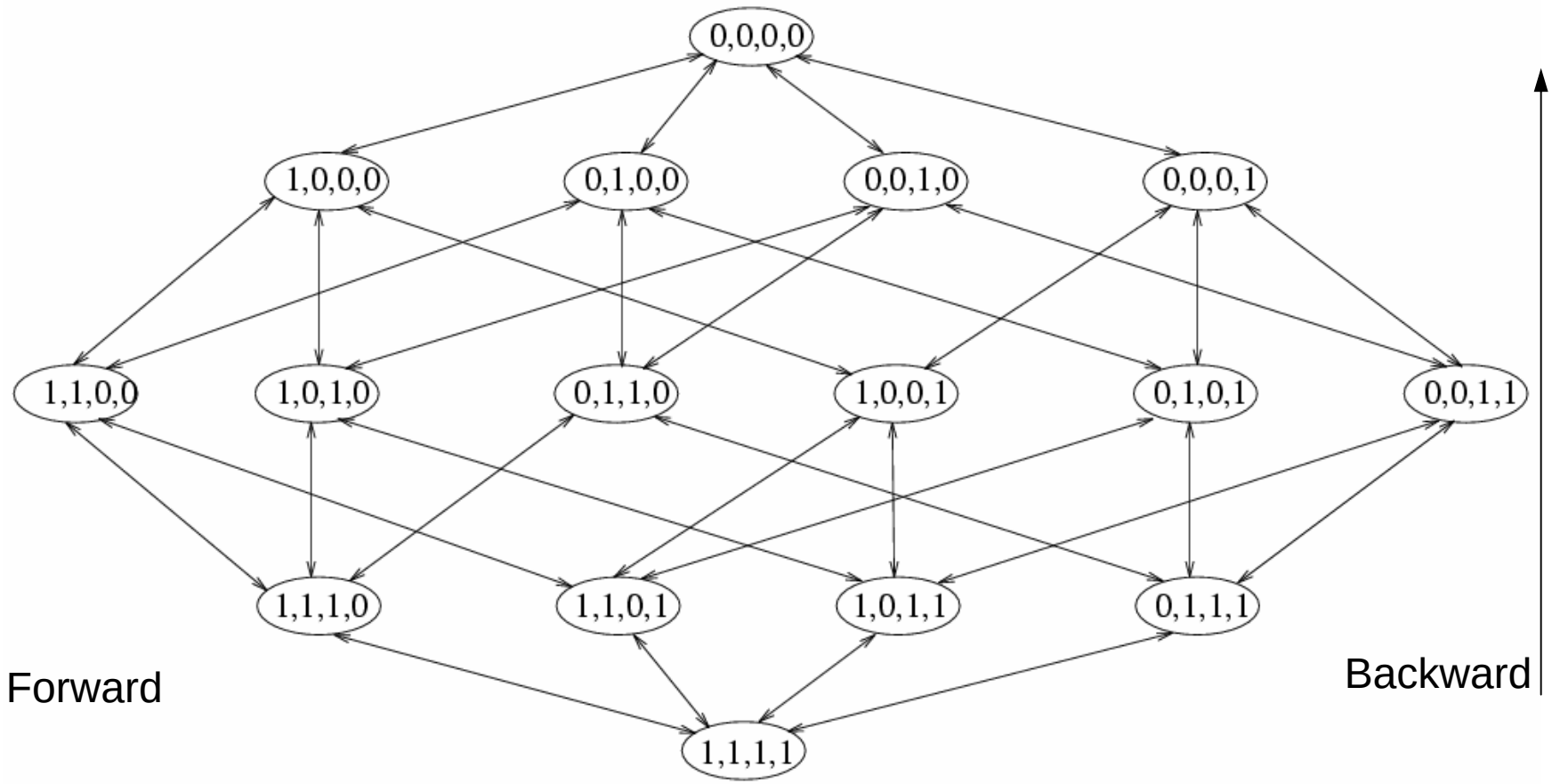
- Data set (whole set)
  - Five Boolean features
  - $C = F_1 \vee F_2$
  - $F_3 = \neg F_2$, $F_5 = \neg F_4$
  - Optimal subset:

    $\{F_1, F_2\}$ or $\{F_1, F_3\}$
- Combinatorial nature of searching for an optimal subset

Ex: How to find the optimal subset?

# A Subset Search Problem

- An example of search space (*Kohavi & John 1997*)



# Different Aspects of Search

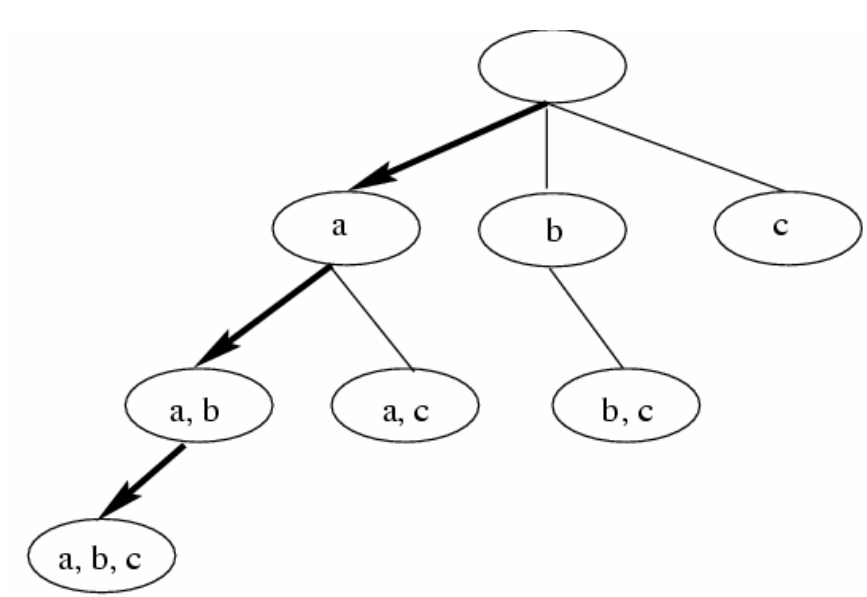
- Search starting points
  - Empty set
  - Full set
  - Random point
- Search directions
  - Sequential forward selection
  - Sequential backward elimination
  - Bidirectional generation
  - Random generation

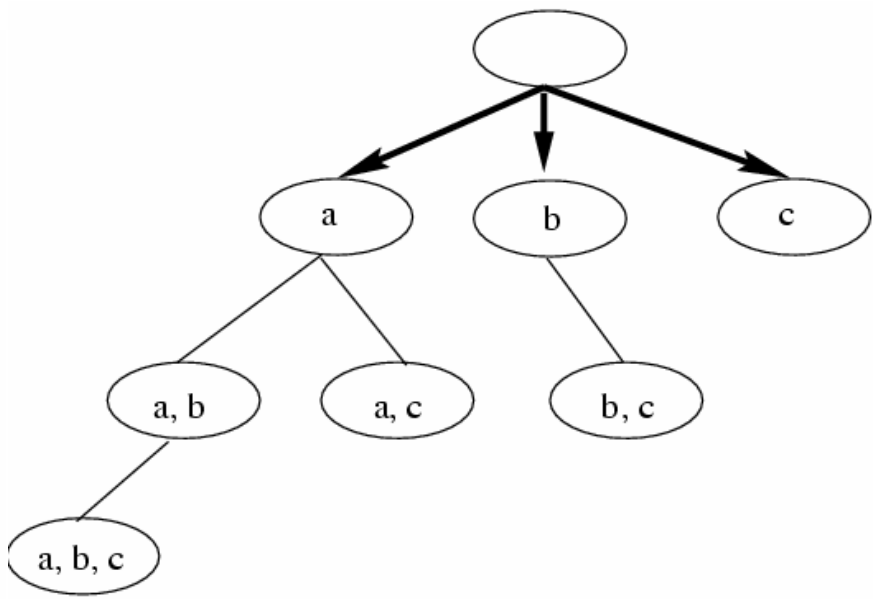
# Different Aspects of Search (Cont'd)

- Search Strategies
  - Exhaustive/complete search
  - Heuristic search
  - Nondeterministic search
- Combining search directions and strategies

| Search Direction | Search Strategy | | |
|---|---|---|---|
| | Complete | Heuristic | Nondeterministic |
| SFG | ✓ | ✓ | ✗ |
| SBG | ✓ | ✓ | ✗ |
| BG | ✓ | ✓ | ✗ |
| RG | ✗ | ✓ | ✓ |

# Illustrations of Search Strategies

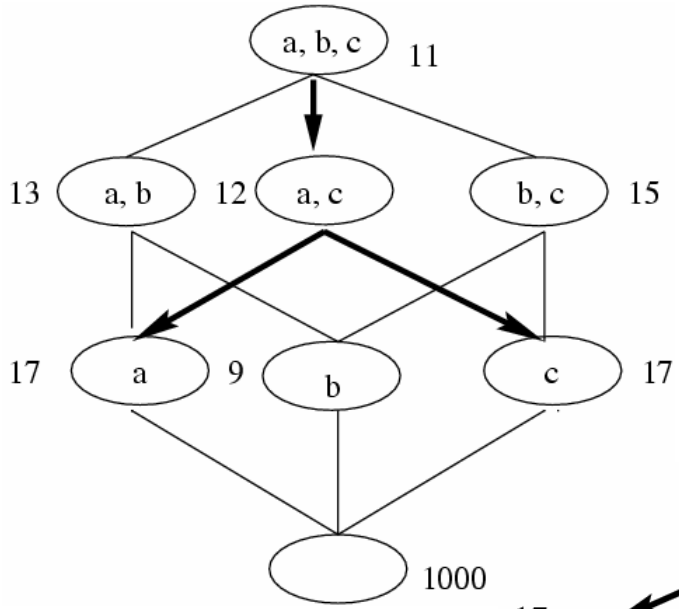


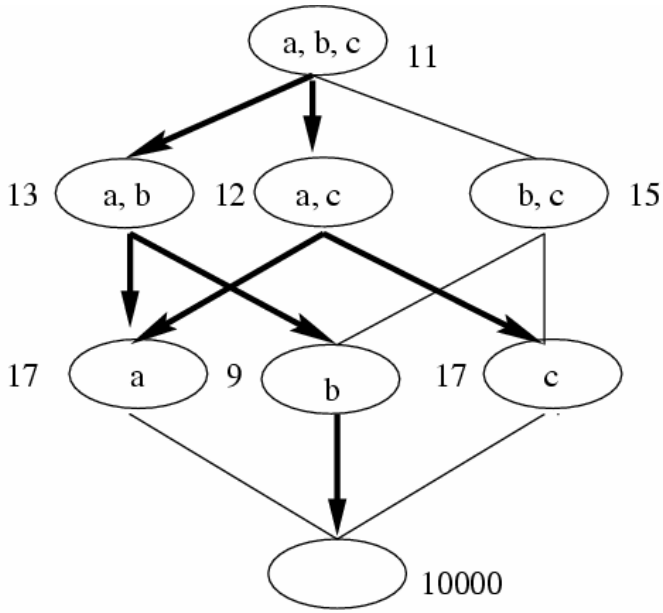

**Depth-first search**

**Breadth-first search**

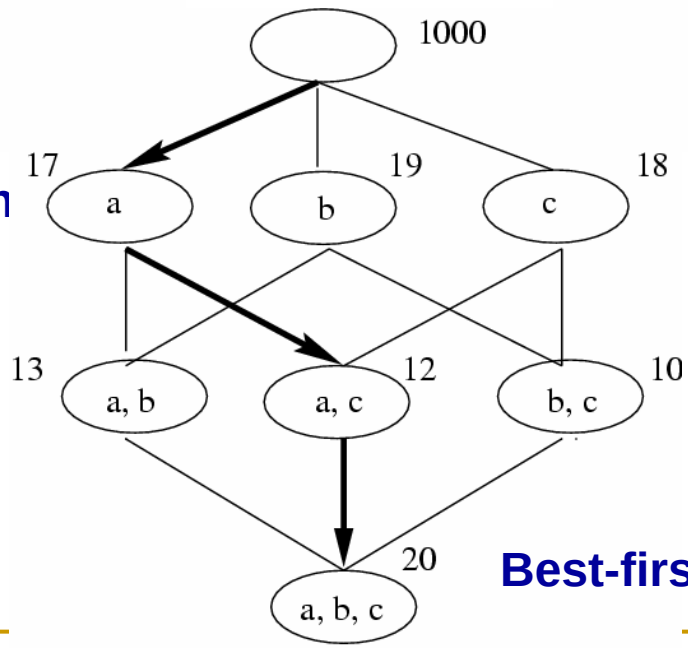
# Illustrations of Search Strategies (Cont'd)




Branch & Bound search




Approx. B & B search




Best-first search

Ex: What are time complexities for different search strategies?

# Feature Ranking

- Weighting and ranking individual features
- Selecting top-ranked ones for feature selection
- Advantages
  - Efficient: $O(N)$ in terms of dimensionality $N$
  - Easy to implement
- Disadvantages
  - Hard to determine the threshold
  - Unable to consider correlation between features

# A General Feature Ranking Algorithm

**Ranking Algorithm**

**Input:** **x** - features, $U$ - measure

**initialize:** list $L = \{\}$ /* $L$ *stores ordered features**/
**for** each feature $x_i$, $i \in \{1, \ldots, N\}$
**begin**
(1) $v_i =$ `compute`$(x_i, U)$
(2) position $x_i$ into $L$ according to $v_i$
**end**

**Output:** $L$ in which the most relevant feature is placed first.

# Evaluation Measures for Ranking and Selecting Features

- The goodness of a feature/feature subset is dependent on measures
- Various measures
  - Information measures (Yu & Liu 2004, Jebara & Jaakkola 2000)
  - Distance measures (Robnik & Kononenko 03, Pudil & Novovicov 98)
  - Dependence measures (Hall 2000, Modrzejewski 1993)
  - Consistency measures (Almuallim & Dietterich 94, Dash & Liu 03)
  - Accuracy measures (Dash & Liu 2000, Kohavi&John 1997)

# Illustrative Data Set (revisit)

| | Hair | Height | Weight | Lotion | Result |
|---|---|---|---|---|---|
| $i_1$ | 1 | 2 | 1 | 0 | 1 |
| $i_2$ | 1 | 3 | 2 | 1 | 0 |
| $i_3$ | 2 | 1 | 2 | 1 | 0 |
| $i_4$ | 1 | 1 | 2 | 0 | 1 |
| $i_5$ | 3 | 2 | 3 | 0 | 1 |
| $i_6$ | 2 | 3 | 3 | 0 | 0 |
| $i_7$ | 2 | 2 | 3 | 0 | 0 |
| $i_8$ | 1 | 1 | 1 | 1 | 0 |

**Sunburn data**

| | Result (Sunburn) | |
|---|---|---|
| | No | Yes |
| P(Result) | 5/8 | 3/8 |
| P(Hair=1\|Result) | 2/5 | 2/3 |
| P(Hair=2\|Result) | 3/5 | 0 |
| P(Hair=3\|Result) | 0 | 1/3 |
| P(Height=1\|Result) | 2/5 | 1/3 |
| P(Height=2\|Result) | 1/5 | 2/3 |
| P(Height=3\|Result) | 2/5 | 0 |
| P(Weight=1\|Result) | 1/5 | 1/3 |
| P(Weight=2\|Result) | 2/5 | 1/3 |
| P(Weight=3\|Result) | 2/5 | 1/3 |
| P(Lotion=0\|Result) | 2/5 | 3/3 |
| P(Lotion=1\|Result) | 3/5 | 0 |

**Priors and class conditional probabilities**

# Information Measures

- Entropy of variable $X$

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i))$$

- Entropy of $X$ after observing $Y$

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j))$$

- Information Gain

$$IG(X|Y) = H(X) - H(X|Y)$$

Ex: Which attribute has the highest info gain for the data of slide 21?

# Consistency Measures

- Consistency measures
  - Trying to find a minimum number of features that separate classes as consistently as the full set can
  - An inconsistency is defined as two instances having the same feature values but different classes
    - E.g., one inconsistency is found between instances i4 and i8 if we just look at the first two columns of the data table (Slide 21)

Ex: Find a smallest subset of features that can maintain consistency for the data of slide 21?

# Accuracy Measures

- Using classification accuracy of a classifier as an evaluation measure
- Factors constraining the choice of measures
  - Classifier being used
  - The speed of building the classifier
- Compared with previous measures
  - Directly aimed to improve accuracy
  - Biased toward the classifier being used
  - More time consuming

Ex: How to obtain reliable accuracy rate?

A brief discussion here, we'll discuss more later.

# Feature Relevance

- Classic definitions (*John et al.,1994*)

  Given $\boldsymbol{F}$, a full set of features, $\boldsymbol{F_i}$, a feature, and

  $\boldsymbol{S_i = F - \{F_i\}}$

**Definition 1 (Strong relevance)** *A feature $F_i$ is strongly relevant iff*

$$\mathbf{P}(C \mid F_i, S_i) \neq \mathbf{P}(C \mid S_i) .$$

**Definition 2 (Weak relevance)** *A feature $F_i$ is weakly relevant iff*

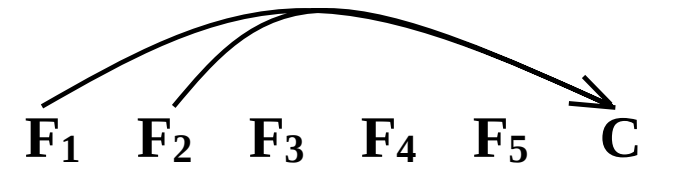
$$\mathbf{P}(C \mid F_i, S_i) = \mathbf{P}(C \mid S_i), \text{ and}$$

$$\exists\, S'_i \subset S_i, \text{ such that } \mathbf{P}(C \mid F_i, S'_i) \neq \mathbf{P}(C \mid S'_i) .$$

**Corollary 1 (Irrelevance)** *A feature $F_i$ is irrelevant iff*

$$\forall\, S'_i \subseteq S_i, \ \mathbf{P}(C \mid F_i, S'_i) = \mathbf{P}(C \mid S'_i) .$$

# An Example for Optimal Subset (revisit)



| $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ | C |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 | 1 |

- Data set (whole set)
  - Five Boolean features
  - $C = F_1 \vee F_2$
  - $F_3 = \neg F_2$, $F_5 = \neg F_4$
  - Optimal subset:

    $\{F_1, F_2\}$ or $\{F_1, F_3\}$
- According to definitions
  - Strongly relevant: $F_1$
  - Weakly relevant: $F_2, F_3$
  - Irrelevant: $F_4$, $F_5$

# Feature Redundancy

- Insufficiency of feature relevance
  - Not able to tell which of weakly relevant features should be selected and which of them removed
- Makov blanket definition (*Koller & Sahami 1996*)

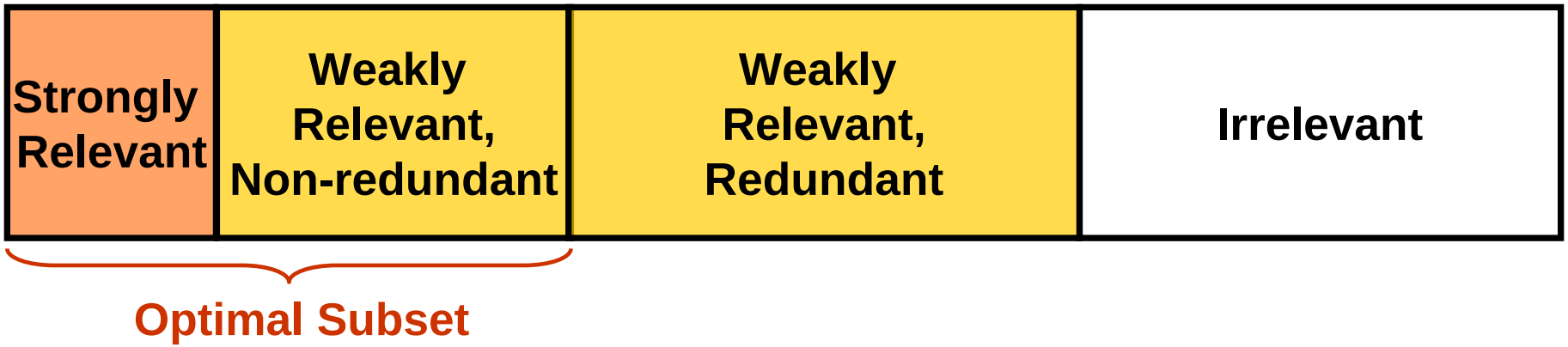
**Definition 3 (Markov blanket)** *Given a feature $F_i$, let $M_i \subset F$ $(F_i \notin M_i)$, $M_i$ is said to be a Markov blanket for $F_i$ iff*

$$\mathbf{P}(F - M_i - \{F_i\},\ C \mid F_i,\ M_i) = \mathbf{P}(F - M_i - \{F_i\},\ C \mid M_i)\ .$$

- Redundant feature definition (*Yu & Liu 2004a*)

**Definition 4 (Redundant feature)** *Let $G$ be the current set of features, a feature is redundant and hence should be removed from $G$ iff it is weakly relevant and has a Markov blanket $M_i$ within $G$ .*

# High-Dimensional Data: Study of Feature Redundancy

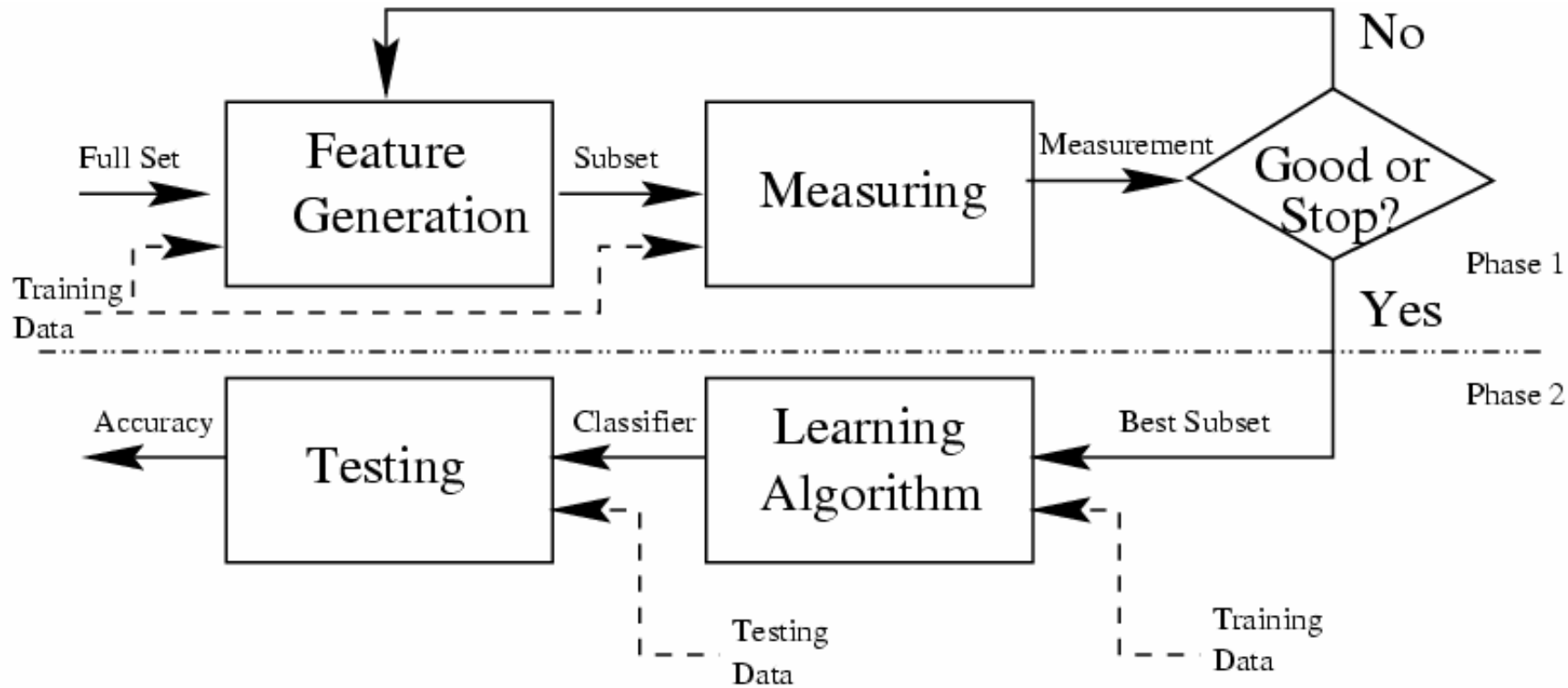



- **Challenges**
  - Thousands of features
  - Many redundant features
- **Goals**
  - Efficiency
  - Effectiveness

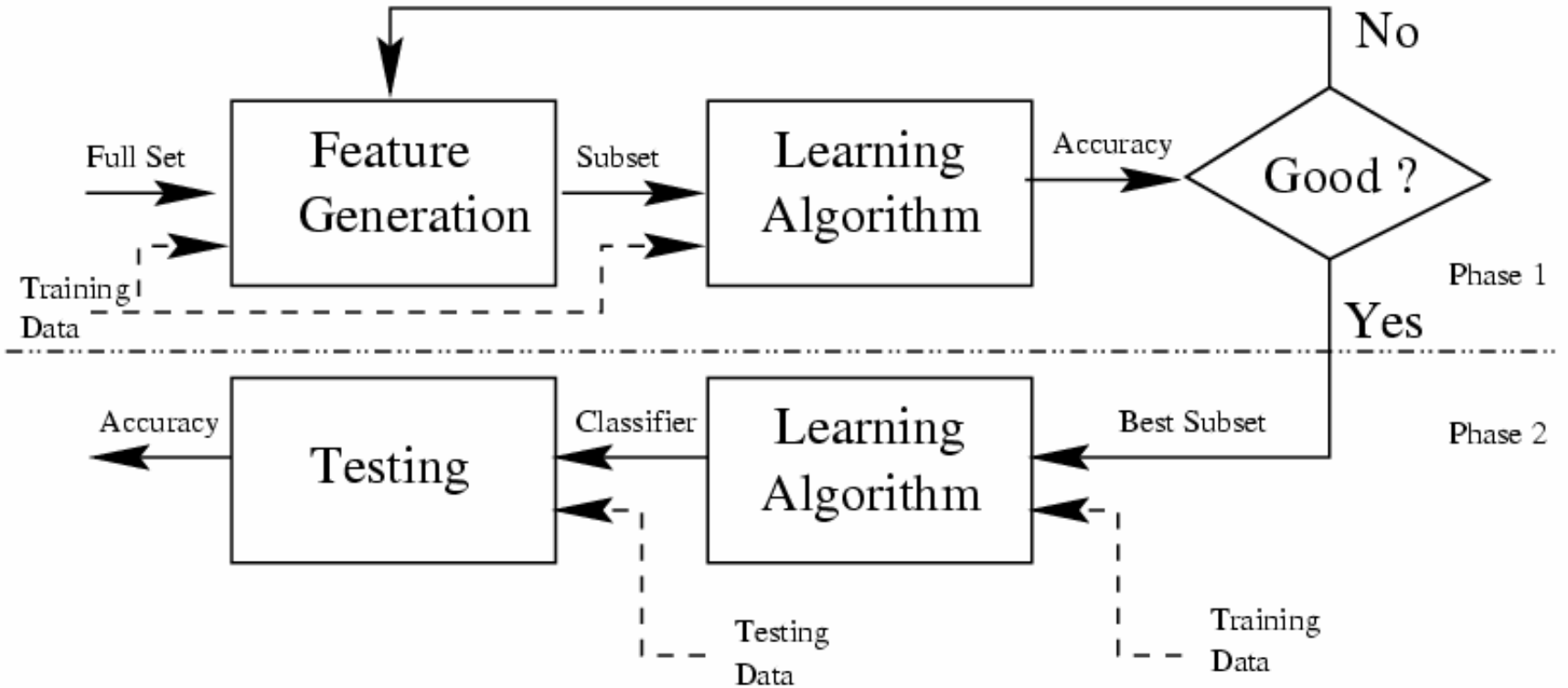
# Models of Feature Selection

- 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

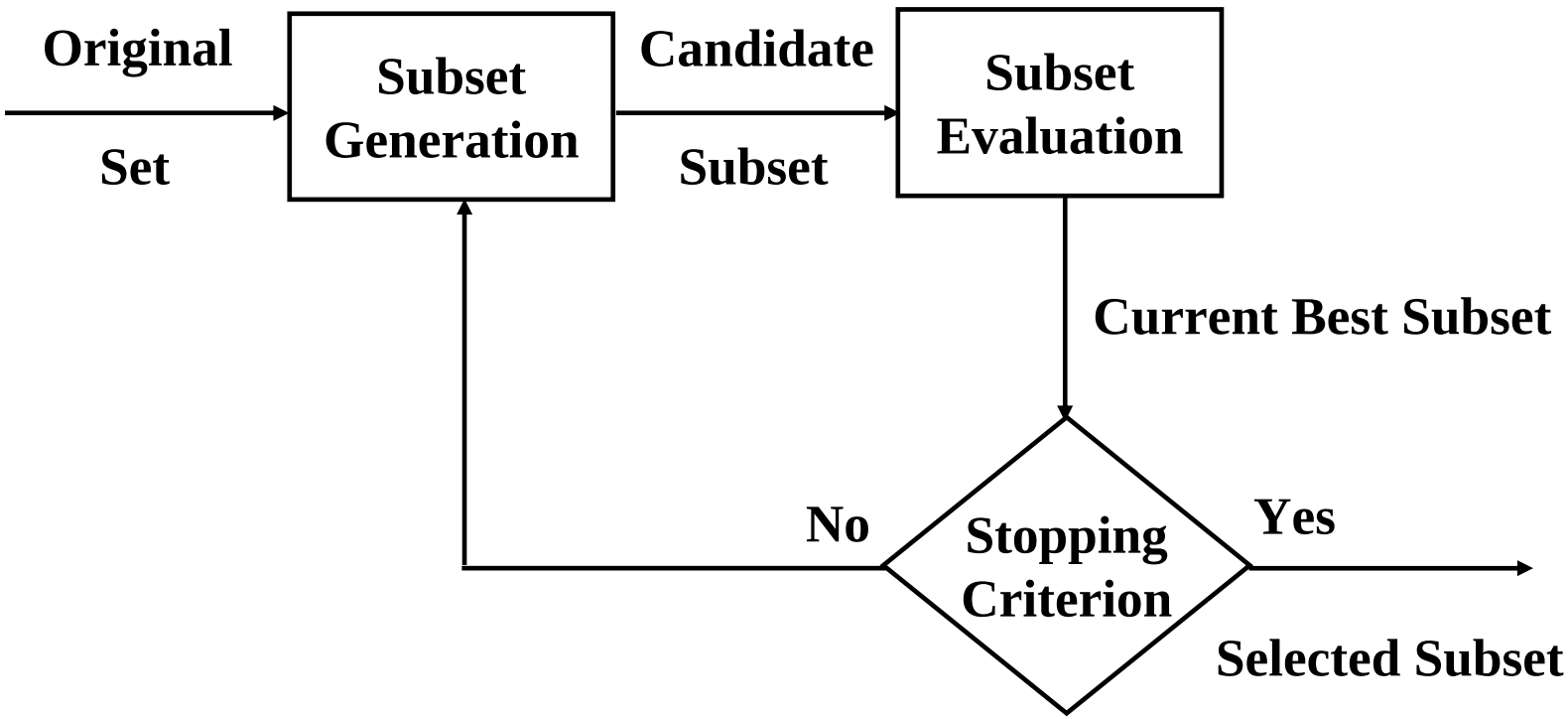
# Filter Model



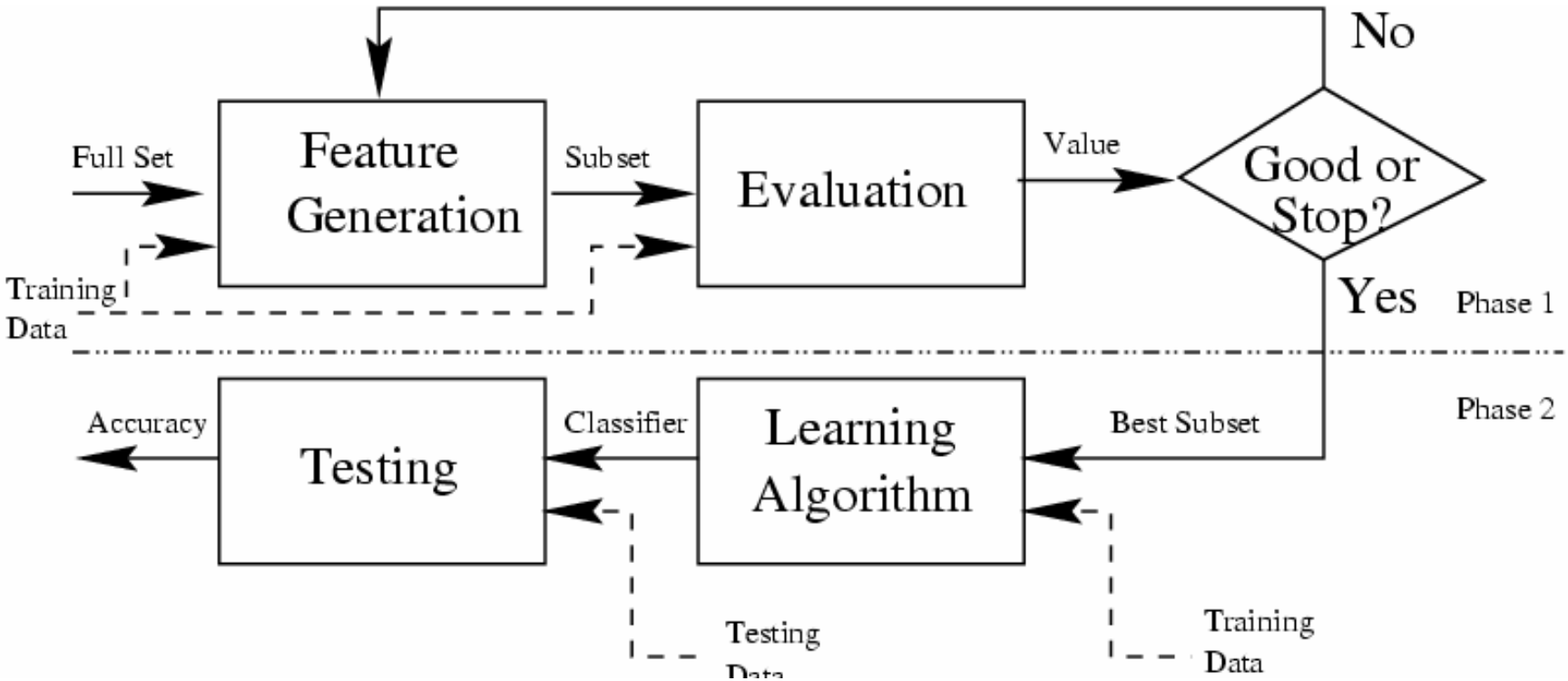
# Wrapper Model



# A Unified View (selection only)



# A Unified View with Selection and Validatation



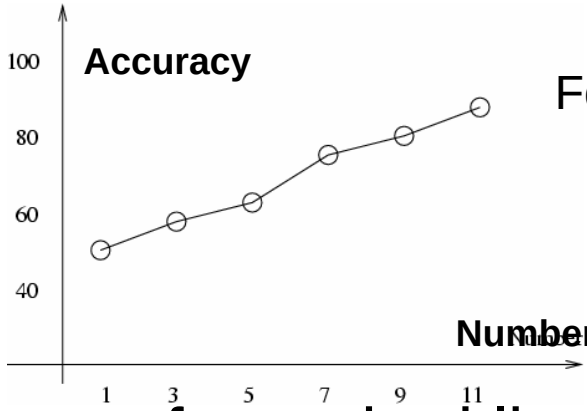
# How to Validate Selection Results

- Direct evaluation (if we know *a priori* …)
  - Often suitable for artificial data sets
  - Based on prior knowledge about data
- Indirect evaluation (if we don't know …)
  - Often suitable for real-world data sets
  - Based on **a)** number of features selected,

    **b)** performance on selected features (e.g., predictive accuracy, goodness of resulting clusters), and **c)** speed

*(Liu & Motoda 1998)*

# Methods for Result Evaluation




For one ranked list

- Learning curves
  - For results in the form of a ranked list of features
- Before-and-after comparison
  - For results in the form of a minimum subset
- Comparison using different classifiers
  - To avoid learning bias of a particular classifier
- Repeating experimental results
  - For non-deterministic results

Ex: What are the proper procedures for evaluation?

# Basics Principles and Algorithms

- Definitions of subset optimality
- Perspectives of feature selection
  - Subset search and feature ranking
  - Evaluation measures
  - Models: filter vs. wrapper
  - Results validation and evaluation
- Representative algorithms for classification
- Selection of algorithms

# Representative Algorithms for Classification

- Filter algorithms
  - Feature ranking algorithms
    - Example: Relief (*Kira & Rendell 1992*)
  - Subset search algorithms
    - Example: consistency-based algorithms
      - Focus (*Almuallim & Dietterich, 1994*)
- Wrapper algorithms
  - Feature ranking algorithms
    - Example
  - Subset search algorithms
    - Example:

# Relief Algorithm

**Relief**

**Input:** **x** - features
$m$ - number of instances sampled
$\tau$ - adjustable relevance threshold

**initialize:** $\mathbf{w} = 0$
**for** $i = 1$ to $m$
**begin**
randomly select an instance $I$
find nearest-hit $H$ and nearest-miss $J$
**for** $j = 1$ to $N$
$\mathbf{w}(j) = \mathbf{w}(j) - \mathtt{diff}(j, I, H)^2/m + \mathtt{diff}(j, I, J)^2/m$
**end**

**Output:** **w** greater than $\tau$

# Relief Algorithm

**Relief**

**Input:** **x** - features
$m$ - number of instances sampled
$\tau$ - adjustable relevance threshold

**initialize:** $\mathbf{w} = 0$
**for** $i = 1$ to $m$
**begin**
randomly select an instance $I$
find nearest-hit $H$ and nearest-miss $J$
**for** $j = 1$ to $N$
$\mathbf{w}(j) = \mathbf{w}(j) - \mathtt{diff}(j, I, H)^2/m + \mathtt{diff}(j, I, J)^2/m$
**end**

**Output:** **w** greater than $\tau$

Ex: What can be candidates for diff()?

Ex: What is its time complexity?

Ex: What are pros and cons of Relief?

# Focus Algorithm

**Focus**

**Input:** $F$ - all features $x$ in data $D$
$U$ - inconsistency rate as evaluation measure

**initialize:** $S = \{\}$

**for** $i = 1$ to $N$
  **for** each subset $S$ of size $i$
    if $\mathrm{Cal}U(S, D) = 0$  /* *CalU (S, D) returns inconsistency**/
      **return** $S$

**Output:** $S$ - a minimum subset that satisfies $U$

# Representative Algorithms for Clustering

- Filter algorithms
  - Example: a filter algorithm based on entropy measure (*Dash et al. 2002*)
- Wrapper algorithms
  - Example: FSSEM – a wrapper algorithm based on EM (expectation maximization) clustering algorithm (*Dy & Brodley 2000*)

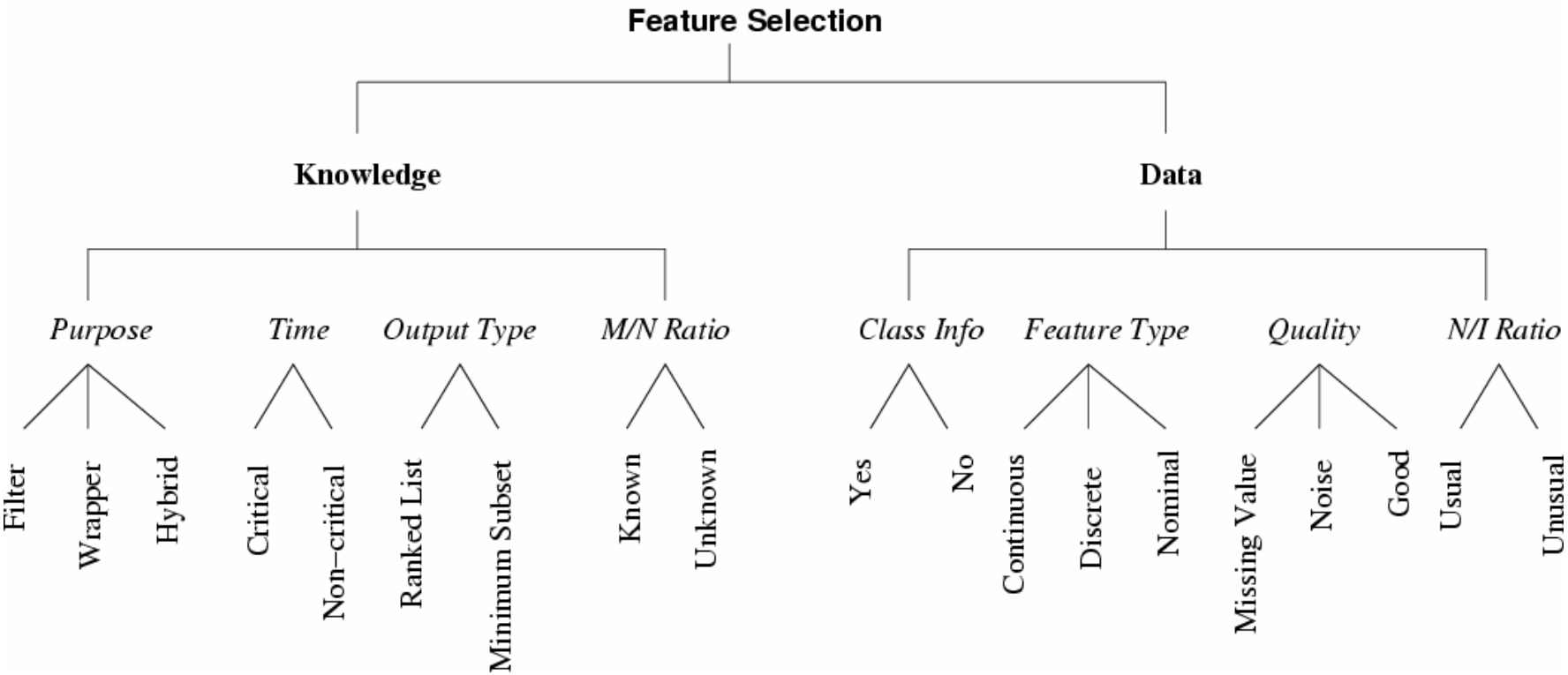
# Categorization of Existing Algorithms

| Evaluation Criteria | | | Search Strategies: Complete – Classification | Complete – Clustering | Sequential – Classification | Sequential – Clustering | Random – Classification | Random – Clustering |
|---|---|---|---|---|---|---|---|---|
| | *Filter* | *Distance* | B&B<br>BFF | | **Relief**<br>ReliefF<br>ReliefS<br>SFS<br>Segen's | | | |
| | | *Information* | MDLM | | DTM<br>Koller's<br>SFG<br>FCBF | **Dash's**<br>SBUD | | |
| | | *Dependency* | Bobrowski's | | CFS<br>RRESET<br>POE+ACC<br>DVMM | Mitra's | | |
| | | *Consistency* | **Focus**<br>**ABB**<br>MIFES1<br>Schlimmer's | | **Set Cover** | | LVI<br>**QBB**<br>**LVF** | |
| | *Wrapper* | *Predictive Accuracy* or *Cluster Goodness* | BS<br>AMB&B<br>FSLC<br>FSBC | | SBS-SLASH<br>WSFG<br>WSBG<br>BDS<br>PQSS<br>RC<br>SS<br>Queiros' | AICC<br>**FSSEM**<br>ELSA | SA<br>RGSS<br>LVW<br>RMHC-PF<br>GA<br>RVE | |
| | *Hybrid* | *Filter+Wrapper* | | | **BBHFS**<br>Xing's | Dash-Liu's | | |
| | | | *Classification* | *Clustering* | *Classification* | *Clustering* | *Classification* | *Clustering* |
| | | | **Data Mining Tasks** | | | | | |

A researher's view

# Guideline for Selecting Algorithms

- A unifying platform




A user's view

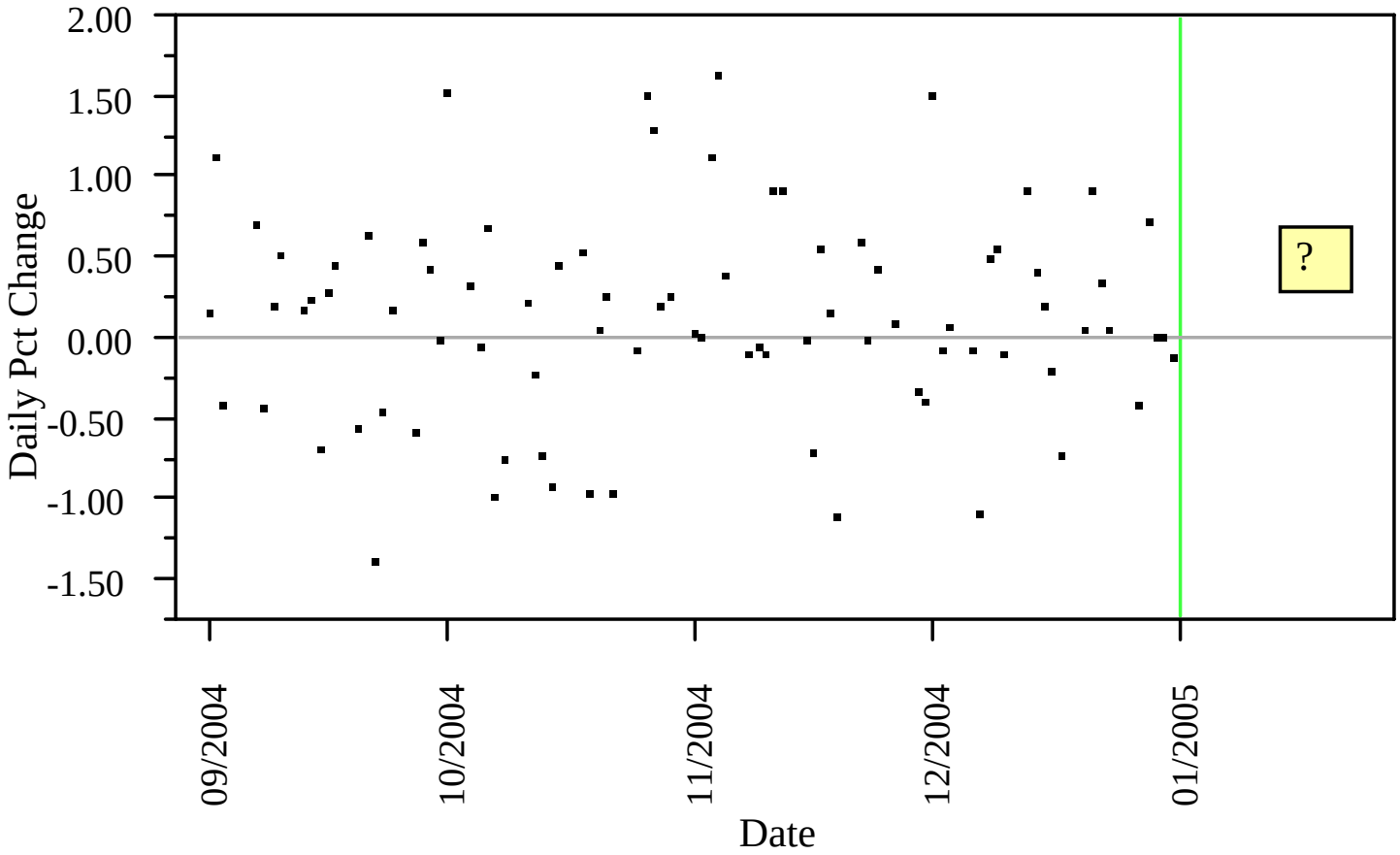
# Model-based Methods: Overview

- Motivating example
- Exploiting probability models
- Two key questions
  - Which features should be considered?
  - Does a feature improve the current classifier?
- Perspective from experience
  - Behavioral and health related applications
    - Credit use, HR management
    - Adverse experiences, drug interactions
  - Many possible features, but few are useful
  - Relatively low signal-to-noise ratio

# Classification Challenge

- Predict whether the stock market is going to increase or decrease tomorrow
- Data
  - Last 4 months of 2004, 85 trading days
  - 12 technical trading rules out of the many
- Model
  - Combine 12 rules to form features
- Test
  - Classify first 97 days of 2005 (through May 20)

# Data record


2.00
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
Daily Pct Change
?
09/2004
10/2004
11/2004
12/2004
01/2005
Date

# Training Results

- Classifier combined given trading rules
- Identifies separating hyperplane that correctly classifies all 85 days in 2004

Predict

| Actual | Count<br>Row % | Better | Worse | |
|---|---|---|---|---|
| | **Better** | 50<br>100.00 | 0<br>0.00 | 50 |
| | **Worse** | 0<br>0.00 | 35<br>100.00 | 35 |
| | | 50 | 35 | 85 |

# Classifier Fails in Test

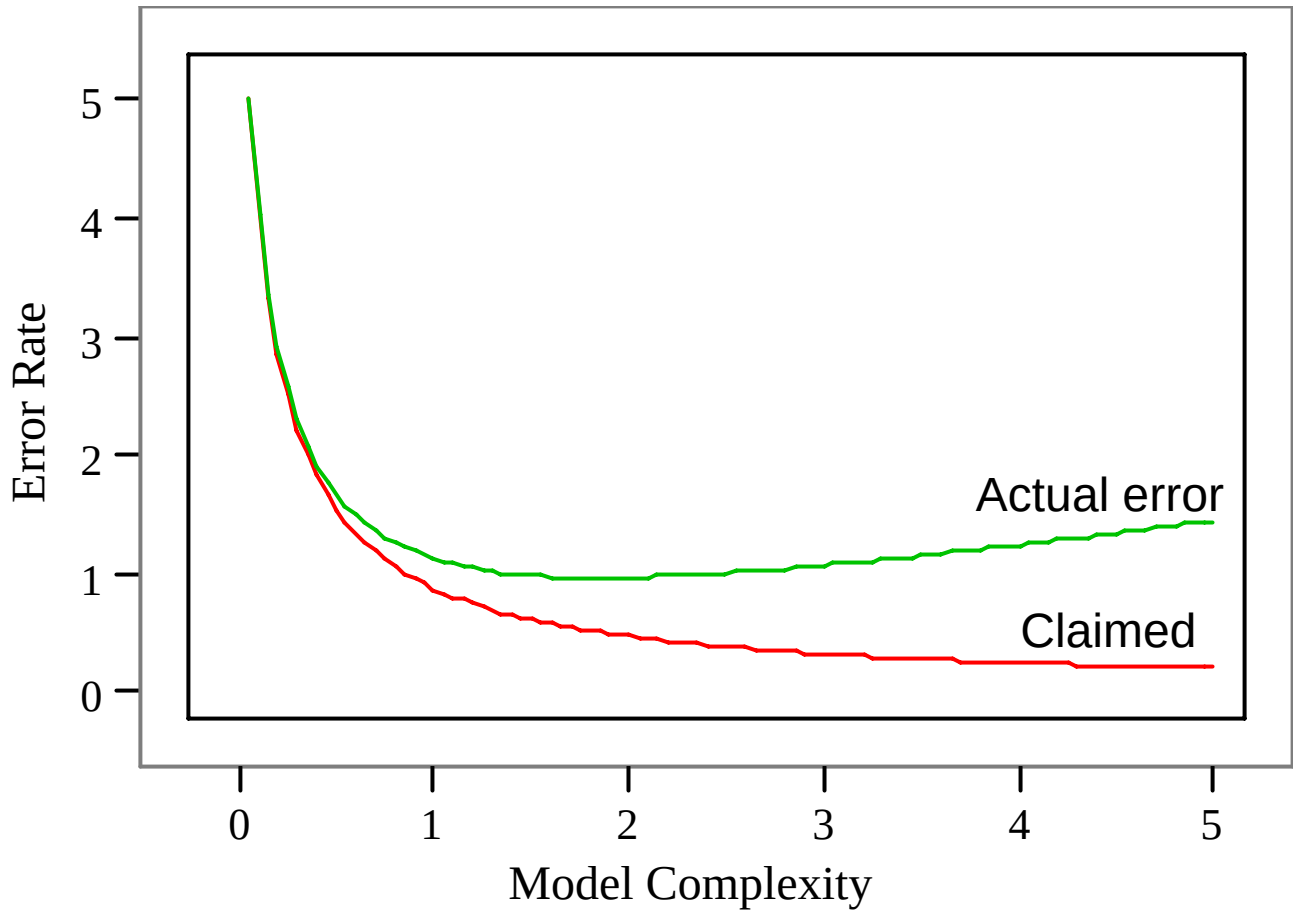
- When the classifier is used in 2005, it fails to beat tossing a coin!

| | | Predicted | | |
|---|---|---|---|---|
| | Count<br>Row % | Better | Worse | |
| Actual | **Better** | 29<br>55.77 | 23<br>44.23 | 52 |
| | **Worse** | 20<br>44.44 | 25<br>55.56 | 45 |
| | | 49 | 48 | 97 |

# Problem: Over-fitting

- Greedy optimization
  - Finds best classifier using training data
- Feature-rich context
  - 12 trading rules
  - + 66 combinations of trading rules
  - + 12 quadratic factors
  - = 90 possible explanations of up/down movement
- Result
  - Combines 30 features into score
  - Classifier encodes artificial pattern that matches random features of trading rules to market.

# Actual versus Claimed Error



# Separating Wheat from Chaff

- When choosing a subset from many candidate features, how can one distinguish
  - Informative, predictive features

  from
  - Coincidental features
- Easy solution: predict test data
- But
  - Willing to use so much data for testing?
  - Is there a way to identify features without sacrificing so much data to testing?

# Testing with Time Series

- Can you afford to wait?
  - Want to model most – all – current data.
  - Can we afford to wait until have enough trading data in order to find out if a rule works?
  - Hold-back test sample is nice for demo, but would this be useful for real financial modeling?
    - Wouldn't you want to use the most recent data?
- Would more historical data help?
  - 90 features, but only 85 training days
  - Do you really want to look back much farther?

# Are you sure this is a poor model?

- What are those predictors?
- Random noise
  - The 12 trading rules are random Gaussian noise
  - A weighted sum of these and their interactions defines a separating hyperplane
  - Hyperplane is perfect in 2004, poor in 2005
- How to avoid the problem of over-fitting?
  - Information-rich times provide too many choices.
  - Hard to separate “luck” from “skill”
  - Possibility of selecting features without CV

# Feature Selection with Models

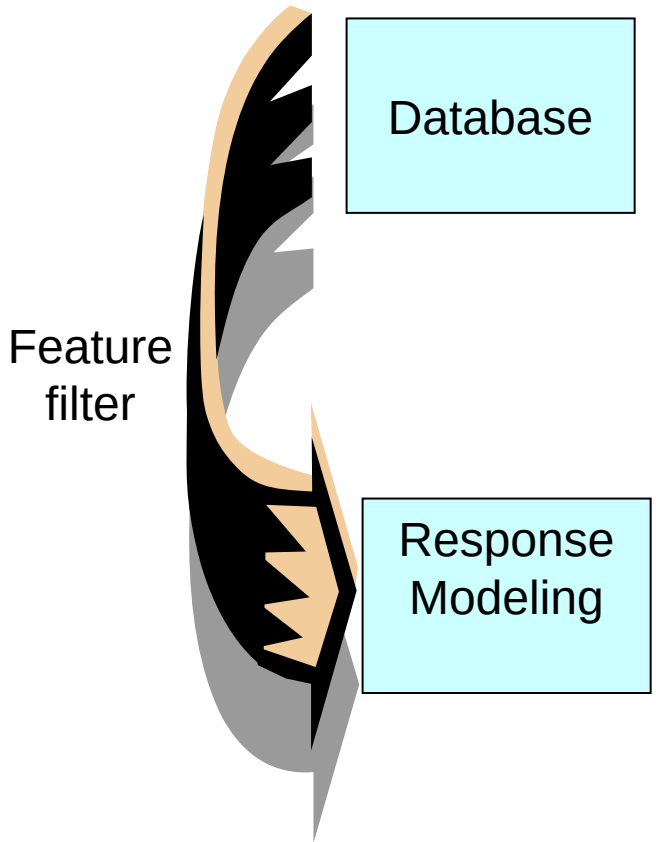
- Wrapper methods
  - Related to a particular model
- When to evaluate features?
  - Joint, at the end of some selection process
    Develop classifier by whatever means, then evaluate the collection of features as a whole
  - Incremental, as each is considered
    Evaluate each feature's contribution, one-at-a time
- How to evaluate features?
  - Computational, using an approach like cross-validation
  - Theoretical, using some type of statistical test that will pick only those that 'guarantee' success

# Advantages of Wrapper Approach

- Wrapper measures contribution of feature within context of the response and algorithm
  - Assess value-added by feature given what is currently used by the algorithm.
  - Compensates for redundancy
    Consider only those features that improve the classification accuracy given what is currently used in the model
- Computation
  - Much can be done off-line, as a hybrid

# Filter/Wrapper Hybrid

- Database holds primitive attributes
- Guided filter “sweeps” effects of current data model from features in database
- Current model guides how filtering is done



# Likelihood-based Feature Evaluation

- Judge contribution of features to learning algorithm based on log-likelihood function

$$L(\beta) = \log P_{\beta}(\text{data})$$

  - Retain only features that improve likelihood by "statistically significant" quantity
  - Requires learning algorithm specify a model that defines probabilities for data
- Questions
  - Which features to try in the model?
  - What is a large improvement?

# Likelihood-based Model

- Assigns *probability* of class membership rather than only a class label to instances
  - Examples
    logistic regression, others in GLIM family
- Log-likelihood is then sum of logs of the probabilities assigned to the cases

$$L_\beta(y_1,\ldots,y_n) = \log P_\beta(y_1) + \ldots + \log P_\beta(y_n)$$

- Assumes independence
  - More general models specify dependence
- Lots of other benefits tag along…

# Advantages of Probability Model

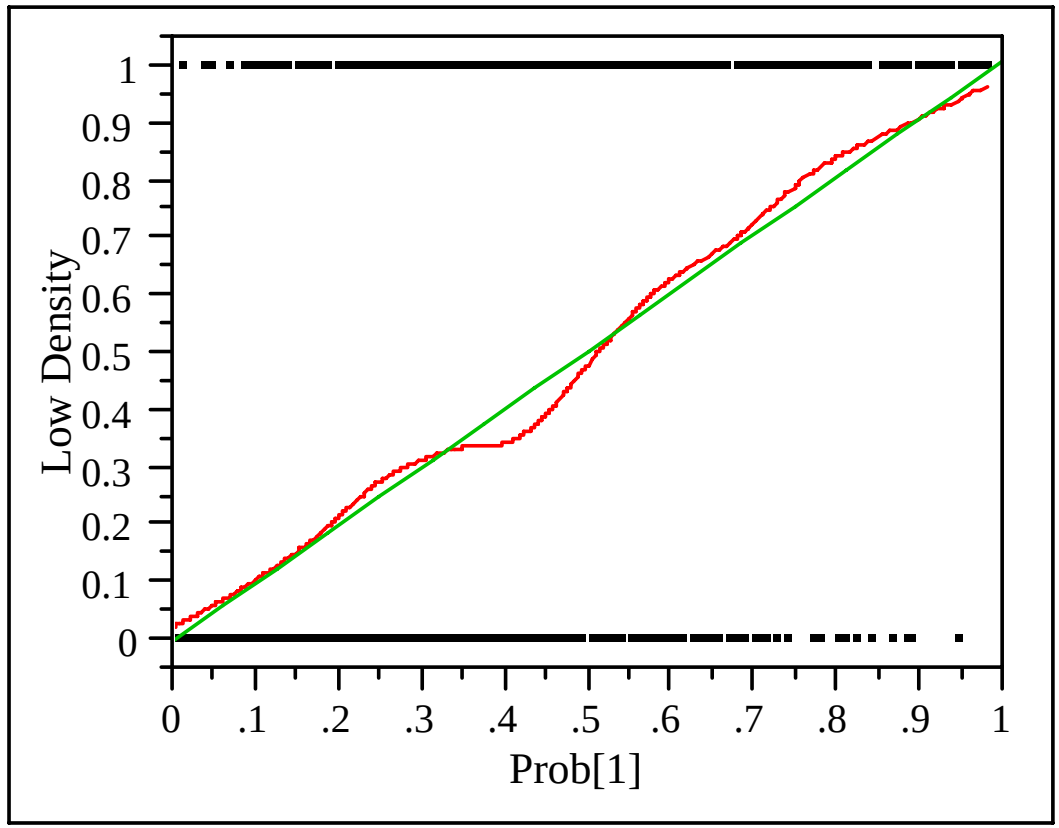
- Cost-based loss function
  - Classify cases based on costs associated with errors of missclassification
- Example
  - If missing a bankrupt customer costs 99 times as much as annoying a good customer, then
    -> Classify as BR if $P_{\beta}(BR) > 0.01$.
- Multiple cut-offs
  - One model handles various thresholds rather than requiring a different model for each

# Probability Model Allows Calibration

- Calibration
  A model is calibrated if its predictions are unbiased in the sense that
  $$\Pr(\text{Class } c \mid \hat{p}) = \hat{p}$$
- Example
  A weather forecaster is calibrated if it rains on 70% of the days that she forecasts “70% chance of rain.”
  - And similarly for other probabilities as well!
- Can always improve uncalibrated classifier

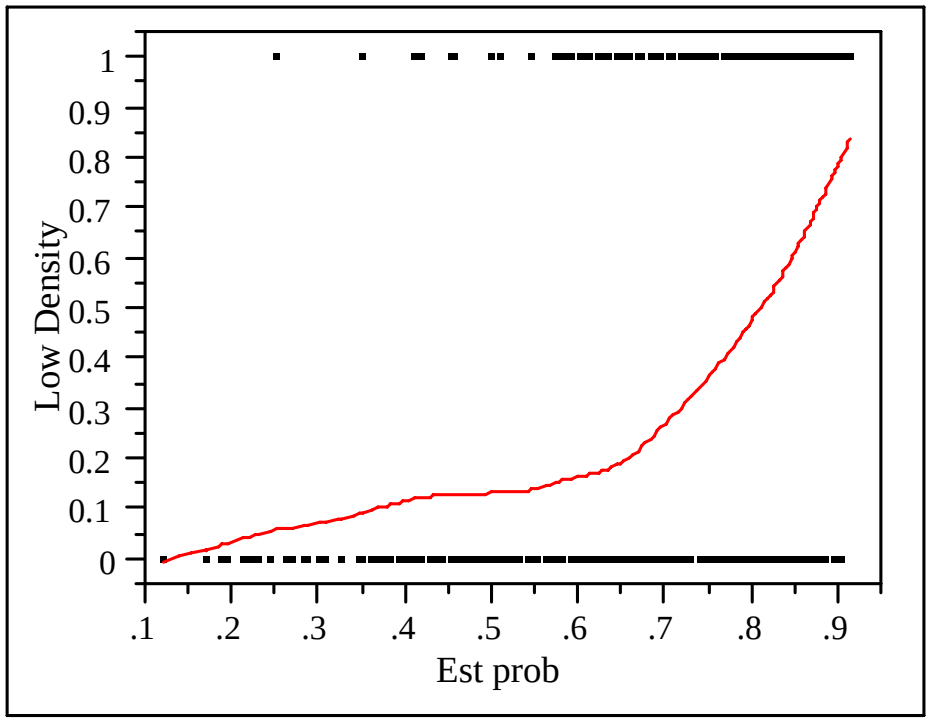
# Example: Well-calibrated

- Risk of low bone mass in women
- Proportions in data align with predicted fraction



# Example: Poorly calibrated

- Estimated probabilities do not match proportions in data
- Eg: Only 10% of those with estimated probability 0.4 display the problem



# Methods for Calibration

- One-dimensional function estimation
  - Wavelets
  - Smoothing splines
  - Polynomials
- Monotone regression
  - Preserve ordering of cases
  - Pooled-adjacent violators
- Predicted values from smoothed function define revised probabilities for classification

# Those Questions…

- Whether you use probability models or other types of learning algorithms, still have to resolve those two questions
  - Which features to consider?
  - Which features that you consider should you use in the model/learning algorithm?

- Start with the easier question…

# Question: Does this feature improve the classifier?

- Source of concern
  Greedy optimization means that added random noise offers an improvement, as in stock example
- Model complexity
  As models become more complex, with more optimization, fit to the training data must improve
- Trade-off
  Complexity added versus gain in fit?
- Probability model offers several equivalent approaches that differ only in motivation

# Description Length

- Represent model as a code for data
  - One part of code describes the data
  - Second part describes model itself
- Information theory
  - Length of code for data given by the log-likelihood, with high probability -> short code
- Trade-off
  - Better fitting models offer short descriptions of the data (ie, good data compression)
  - More complex models require more elaborate, longer descriptions

# Minimum Description Length (MDL)

- Add features to the model so long as the overall description length decreases
- Description length is the log-likelihood plus the length of the model description

$$DL(data) = L_{\beta}(data) + L(\beta)$$

- Various methods for describing the model lead to different criteria for judging models

# MDL Criterion

- Principle
  Add feature to model if the change in the description length for data is enough to "pay for" length needed to describe model
- Model indexed by $\beta$ is better than model indexed by $\theta$ if

$$L_{\beta}(\text{data}) + L(\beta) < L_{\theta}(\text{data}) + L(\theta)$$

- Change in description length for the model depends on how model is encoded
  - Choice for model form depends on "prior" beliefs

# Equivalence among Methods

| *Method* | *Principle* |
|---|---|
| Minimum description length | Shortest overall code length for data and model |
| Penalized likelihood | Regularization to introduce stability among estimated parameters |
| Bayesian model selection | Prior distribution on model parameters |

# Commonly used Rules

| Δ log likelihood | P-value | Name | Author |
|---|---|---|---|
| 2 | 0.16 | AIC, leave-1-out cross-validation | Akaike, Mallows 1973 |
| 4 | 0.05 | Classical test | |
| Log n | – | MDL, BIC | Rissanen, Schwarz 1978 |
| 2 Log m | 1/m | RIC | Donoho&Johnstone Foster&George 1994 |

# Role of p-value

- P-value measures "probability" of feature effect under the assumption $H_0$ that the feature adds no value
  - Small p-value implies either a miracle or feature adds predictive value to classifier
- Typically computed under normal-theory model for sampling variation (CLT)
- Variations allow adjustments for lack of normality in problems with rare events
  - Bennett bounds (Foster&Stine 2004)

# Trends in Methods

- Early approaches
  - Get an unbiased estimate of the out-of-sample errors of a model (AIC)
  - Pick the model with the smallest estimated out-of-sample error
- Problem
  - Selection bias happens when have many similar models, choice wins by chance alone
- More recent methods
  - Penalize for size of search space (RIC)

# Role for Cross-Validation?

- Criterion methods "do it all"
  - Avoid cross validation by picking features endogenously without reserving hold-back sample
- Avoiding cross-validation necessary when
  - Time series, forecasting out-of-sample
  - Too little data to sacrifice to cross-validation
  - Computations too slow
  - Too many models to compare

# Question: Which features to consider?

- Penalty methods provide a scale of "goodness" to grade features, but
- How does one move from a model with certain features to the next?
  - Classical gradient-type methods (ie stepwise)
  - Regularization approach
  - Exogenous ordering

# Regularization Approach

- Common regularization adds quadratic penalty to log likelihood (eg, ridge regression)

$$\log P_\beta(\text{data}) + \lambda \sum \beta_\iota^2$$

  - Shrinks toward zero, but results in model of high dimension.
- Recent interest in L1 penalty (eg, lasso)

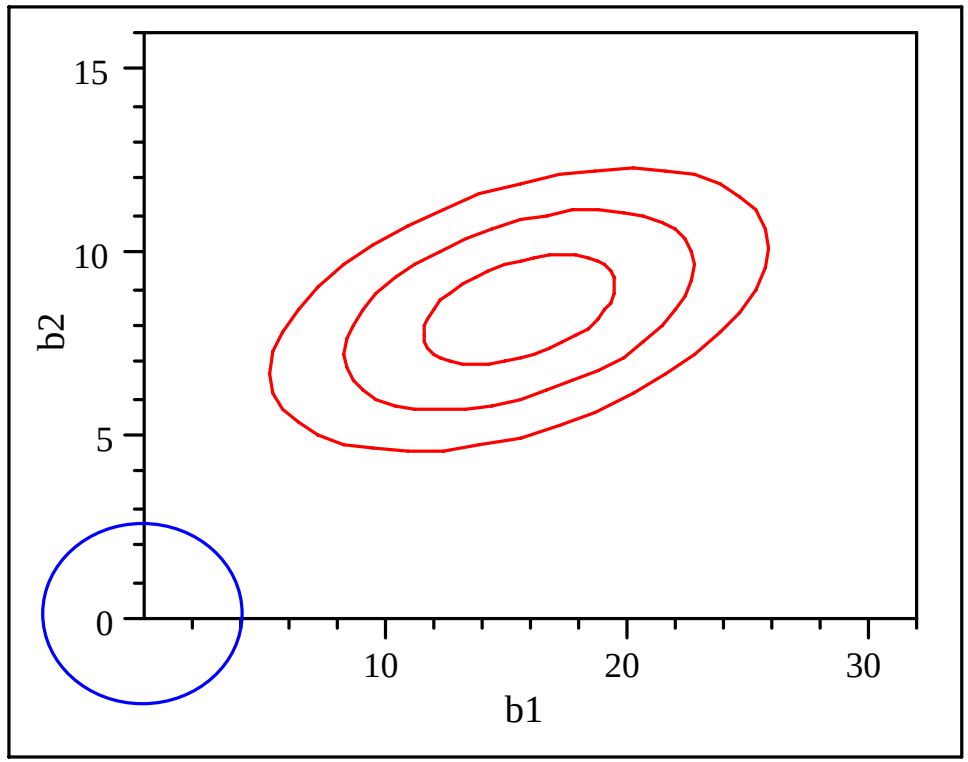
$$\log P_\beta(\text{data}) + \lambda \sum |\beta_\iota|$$

  - L1 produces selection rather shrinkage
  - Compromises (eg elastic net) blend the two approaches

# Picture Explains the Differences

- L2 penalty

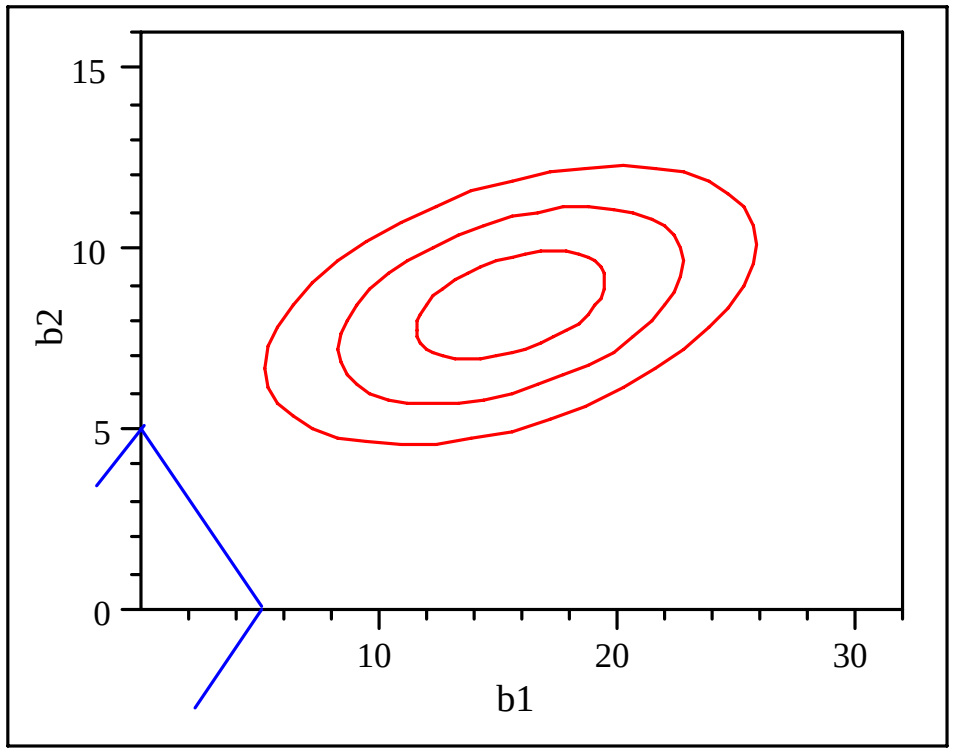
  Rounded boundary of constraint does not zero any parameter



# Picture Explains the Differences

- L1 penalty

  Sharp corners along axes zero parameters



# Follow Regularization Path

- Generate sequence of models
  - Vary the level of the constraint $\lambda$ in the regularized log-likelihood

    $$\log P_{\beta}(\text{data}) + \lambda \sum |\beta_{\iota}|$$

  - Sequence defines "regularization path"
- Role for cross-validation
  - L1 penalty produces a sequence of models
  - Cross validation picks models along this regularization path.
  - Rather than considering all possible, use L1 penalty to define sequence of interesting models

# Discussion of Regularization

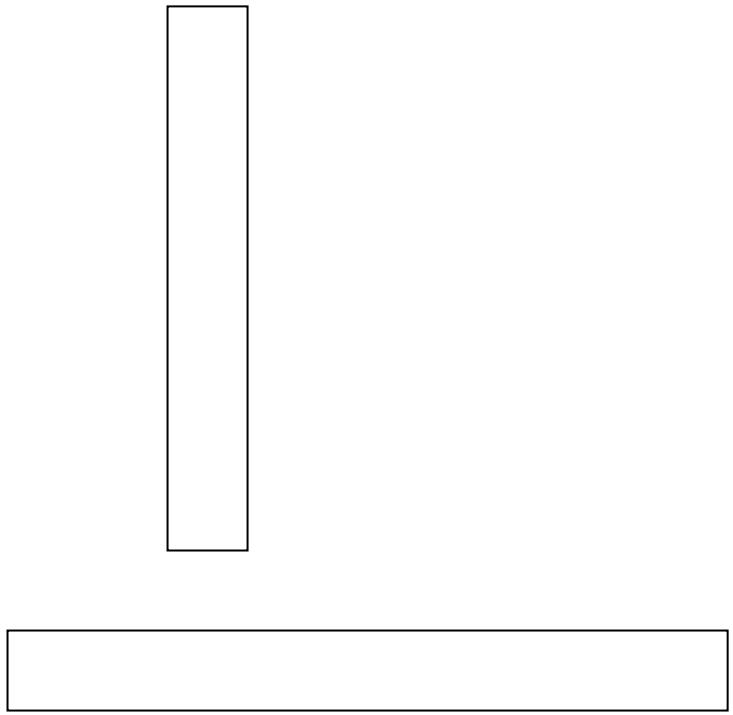
- Scales
  - Requires a common scaling, weighting
- Computation
  - Trust-region approach to optimization
  - Linear paths in SVM
- Connections
  - L2 penalty equivalent to Normal prior
  - L1 penalty equivalent to Laplace prior
  - Shapes of joint distributions differ

# Model-based Methods: Summary

- Have focused on avoiding over-fitting in problems with many features
- Key questions
  - Which features to consider?
    - Regularization paths
    - Greedy, gradient methods
    - Substantive covered later
  - Which features to keep?
    - Role of p-values, probability models in assessment
    - Statistical methods that avoid cross validation

# Dealing with Wide Data (III)

- Handling large-sized data (a large number of instances)
- Handling high-dimensional data (a large number of features)



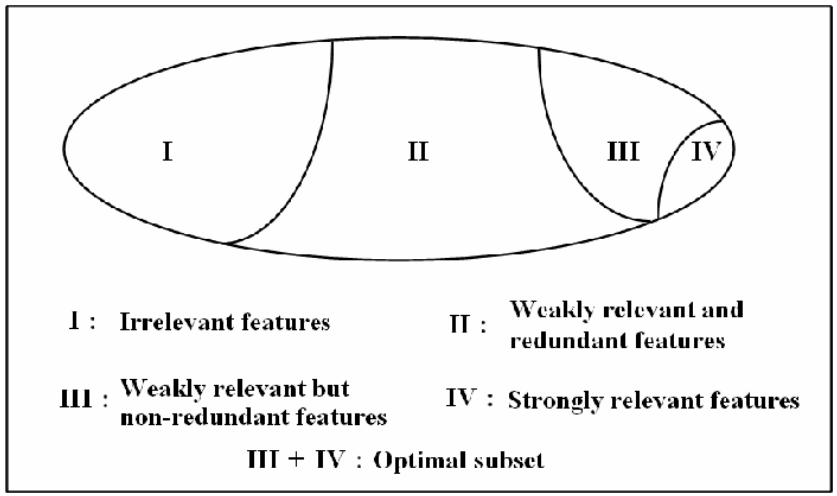
# Handling High-dimensional Data

- High-dimensional data
  - As in gene expression microarray analysis, text categorization, …
  - With hundreds to tens of thousands of features
  - With many irrelevant and redundant features
- Some research efforts
  - Redundancy based feature selection
  - Feature selection for text classification

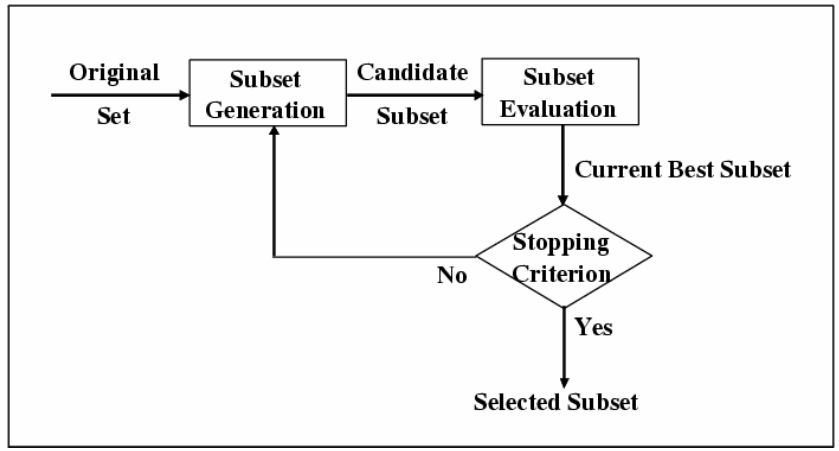
# Need for Relevance-based Feature Selection

- Individual feature evaluation
  - Focusing on identifying relevant features without handling feature redundancy, roughly $O(N)$
- Feature subset evaluation
  - Relying on minimum feature subset heuristics to implicitly handling redundancy while pursuing relevant features, at least $O(N^2)$
- Effectiveness and efficiency
  - Able to handle both irrelevant and redundant features
  - Less costly than existing subset evaluation methods

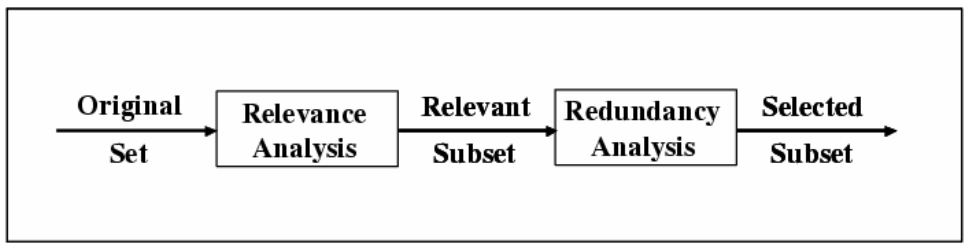
# Our Solution – A New Framework of Feature Selection




A view of feature relevance and redundancy




A traditional framework of feature selection




A new framework of feature selection

# An Approximation Method (Yu & Liu 2004)

- Need for approximation methods
  - Searching for an optimal subset is combinatorial
  - Over-searching on training data can cause over-fitting for wide data
- Two steps of approximation
  - To approximately find the set of relevant features
  - To approximately determine feature redundancy among relevant features

# Evaluation Measure

- Entropy-based measure
  - Symmetrical Uncertainty ($SU$)
    - Symmetry for $X$ and $Y$
    - Range [0, 1]

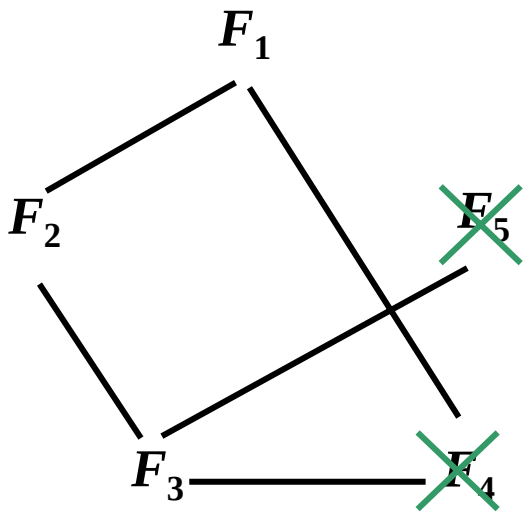
$$SU(X,Y) = 2\frac{H(X) - H(X|Y)}{H(X) + H(Y)}$$

- Two types of correlation by $SU$ value
  - $C$-correlation (feature $F_i$ and class $C$): $SU_{i,c}$
  - $F$-correlation (feature $F_i$ and $F_j$): $SU_{i,j}$

# Approximation of Relevant Features

- Aiming to achieve high efficiency
  - Calculate $C$-correlation for each feature
  - Heuristically decide a feature $F_i$ to be relevant if it is highly correlated with the class $C$, i.e., $SU_{i,c} \geqslant \delta$
- Selected relevant features are subject to redundancy analysis

# Determining Redundancy

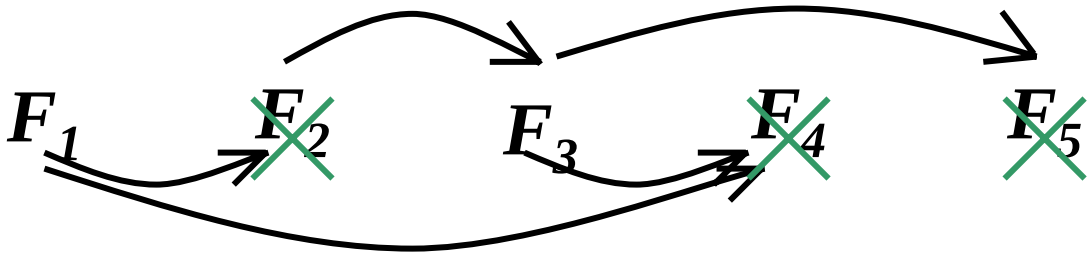



- Hard to decide redundancy
  - Redundancy criterion
  - Which one to keep
- Approximate redundancy criterion

  $F_j$ is redundant to $F_i$ *iff*

  $SU(F_i, C) \geqslant SU(F_j, C)$ and $SU(F_i, F_j) \geqslant SU(F_j, C)$


- Predominant feature: not redundant to any feature in the current set



# Algorithm

- Fast Correlation-Based Filter (FCBF)
  - Calculate $SU$ value for each feature, order them, select relevant features based on a threshold
  - Start with the first feature (as a predominant feature) to eliminate all features for which it forms an approximate redundant cover
  - Continue with the next remaining feature until the end of list
- Efficiency and effectiveness
  - $O(N)$ to remove irrelevant features
  - On average $O(N\log N)$ to remove redundant features

# 10 UCI Bench-mark Data Sets

| Title | Features | Instances | Classes |
|---|---|---|---|
| Lung-cancer | 57 | 32 | 3 |
| Promoters | 59 | 106 | 2 |
| Splice | 62 | 3190 | 3 |
| USCensus90 | 68 | 9338 | 3 |
| CoIL2000 | 86 | 5822 | 2 |
| Chemical | 151 | 936 | 3 |
| Musk2 | 169 | 6598 | 2 |
| Arrhythmia | 280 | 452 | 16 |
| Isolet | 618 | 1560 | 26 |
| Multi-features | 650 | 2000 | 10 |

# Speed

| Title | Running Time | | | |
|---|---|---|---|---|
| | FCBF | CorrSF | ReliefF | ConsSF |
| Lung-cancer | 20 | 50 | 50 | 110 |
| Promoters | 20 | 50 | 100 | 190 |
| Splice | 200 | 961 | 2343 | 34920 |
| USCensus90 | 541 | 932 | 7601 | 161121 |
| CoIL2000 | 470 | 3756 | 7751 | 341231 |
| Chemical | 121 | 450 | 2234 | 14000 |
| Musk2 | 971 | 8903 | 18066 | 175453 |
| Arrhythmia | 151 | 2002 | 2233 | 31235 |
| Isolet | 3174 | 177986 | 17025 | 203973 |
| Multi-Features | 4286 | 125190 | 21711 | 133932 |
| Average | 995 | 32028 | 7911 | 109617 |

# Degree of Dimensionality Reduction

| Title | # Selected Features | | | |
|---|---|---|---|---|
| | FCBF | CorrSF | ReliefF | ConsSF |
| Lung-cancer | 5 | 8 | 5 | 4 |
| Promoters | 4 | 4 | 4 | 4 |
| Splice | 6 | 6 | 11 | 10 |
| USCensus90 | 2 | 1 | 2 | 13 |
| CoIL2000 | 3 | 10 | 12 | 29 |
| Chemical | 4 | 7 | 7 | 11 |
| Musk2 | 2 | 10 | 2 | 11 |
| Arrhythmia | 6 | 25 | 25 | 24 |
| Isolet | 23 | 137 | 23 | 11 |
| Multi-Features | 14 | 87 | 14 | 7 |
| Average | 7 | 30 | 11 | 12 |

# Predictive Accuracy (C4.5)

| Title | Full Set | FCBF | CorrSF | ReliefF | ConsSF |
|---|---|---|---|---|---|
| Lung-cancer | 80.83 ±22.92 | 87.50 ±16.32 | 84.17 ±16.87 | 80.83 ±22.92 | 84.17 ±16.87 |
| Promoters | 86.91 ±6.45 | 87.73 ±6.55 | 87.73 ±6.55 | 89.64 ±5.47 | 84.00 ±6.15 |
| Splice | 94.14 ±1.57 | 93.48 ±2.20 | 93.48 ±2.20 | 89.25 ±1.94 | 93.92 ±1.53 |
| USCensus90 | 98.27 ±0.19 | 98.08 ±0.22 | 97.95 ±0.15 | 98.08 ±0.22 | 98.22 ±0.30 |
| CoIL2000 | 93.97 ±0.21 | 94.02 ±0.07 | 94.02 ±0.07 | 94.02 ±0.07 | 93.99 ±0.20 |
| Chemical | 94.65 ±2.03 | 95.51 ±2.31 | 96.47 ±2.15 | 93.48 ±1.79 | 95.72 ±2.09 |
| Musk2 | 96.79 ±0.81 | 91.33 ±0.51 | 95.56 ±0.73 | 94.62 ±0.92 | 95.38 ±0.75 |
| Arrhythmia | 67.25 ±3.68 | 72.79 ±6.30 | 68.58 ±7.41 | 65.90 ±8.23 | 67.48 ±4.49 |
| Isolet | 79.10 ±2.79 | 75.77 ±4.07 | 80.70 ±4.94 | 52.44 ±3.61 | 69.23 ±4.53 |
| Multi-Features | 94.30 ±1.49 | 95.06 ±0.86 | 94.95 ±0.96 | 80.45 ±2.41 | 90.80 ±1.75 |
| Average | 88.62 ±9.99 | 89.13 ±8.52 | 89.36 ±9.24 | 83.87 ±14.56 | 87.29 ±11.04 |

# Predictive Accuracy (NBC)

| Title | Full Set | FCBF | CorrSF | ReliefF | ConsSF |
|---|---|---|---|---|---|
| Lung-cancer | 80.00 ±23.31 | 90.00 ±16.10 | 90.00 ±16.10 | 80.83 ±22.92 | 86.67 ±17.21 |
| Promoters | 90.45 ±7.94 | 94.45 ±8.83 | 94.45 ±8.83 | 87.82 ±10.99 | 92.64 ±7.20 |
| Splice | 95.33 ±0.88 | 93.60 ±1.74 | 93.60 ±1.74 | 88.40 ±1.97 | 94.48 ±1.39 |
| USCensus90 | 93.38 ±0.90 | 97.93 ±0.16 | 97.95 ±0.15 | 97.93 ±0.16 | 97.87 ±0.26 |
| CoIL2000 | 79.03 ±2.08 | 93.94 ±0.21 | 92.94 ±0.80 | 93.58 ±0.43 | 83.18 ±1.94 |
| Chemical | 60.79 ±5.98 | 72.11 ±2.51 | 70.72 ±4.20 | 78.20 ±3.58 | 67.20 ±2.51 |
| Musk2 | 84.69 ±2.01 | 84.59 ±0.07 | 64.85 ±2.09 | 84.59 ±0.07 | 83.56 ±1.05 |
| Arrhythmia | 60.61 ±3.32 | 66.61 ±5.89 | 68.80 ±4.22 | 66.81 ±3.62 | 68.60 ±7.64 |
| Isolet | 83.72 ±2.38 | 80.06 ±2.52 | 86.28 ±2.14 | 52.37 ±3.17 | 71.67 ±3.08 |
| Multi-Features | 93.95 ±1.50 | 95.95 ±1.06 | 96.15 ±0.94 | 76.05 ±3.26 | 93.75 ±1.95 |
| Average | 82.20 ±12.70 | 86.92 ±10.79 | 85.57 ±12.53 | 80.66 ±13.38 | 83.96 ±11.31 |

# Summary of Results on UCI Data

- Average results on 10 data sets

| | Full Set | FCBF | CorrSF | ReliefF | ConsSF |
|---|---|---|---|---|---|
| Running time (ms) | | 995 | 32028 | 7911 | 109617 |
| # Features | 220 | 7 | 30 | 11 | 12 |
| Accuracy on C4.5 | 88.62±9.99 | 89.13±8.52 | 89.36±9.24 | 83.87±14.56 | 87.29±11.04 |
| Accuracy on NBC | 82.20±12.70 | 86.92±10.79 | 85.57±12.53 | 80.66±13.38 | 83.96±11.31 |

- FCBF can achieve the following
  - Faster speed
  - Higher degree of dimensionality reduction
  - Improved classification accuracy

# Feature Selection in Text Categorization

- A comparative study in (*Yang & Pederson 1997*)
  - 5 metrics evaluated and compared
    - Document Frequency (DF), Information Gain (IG), Mutual Information (MU), $\chi^2$ statistics (CHI), Term Strength (TS)
    - IG and CHI performed the best
  - Improved classification accuracy of *k*-NN achieved after removal of up to 98% unique terms by IG
- Another study in (*Forman 2003*)
  - 12 metrics evaluated on 229 categorization problems
  - A new metric, Bi-Normal Separation, outperformed others and improved accuracy of SVMs

# Feature Selection in Text Categorization

## Applications

- Web pages
  - Recommending, Yahoo-like classification
- Newsgroup Messages
  - Recommending, spam filtering
- News articles
  - Personalized newspaper
- Email messages
  - Routing, Prioritizing, Spam filtering

# Documents Representation

- Vector Space Model
- A term can be
  - Word, or phrase (syntactic phrase, N-grams)
- Weight of term
  - Boolean, Term Frequency (tf)
  - tf*idf (Inverse document frequency)
- Properties
  - High Dimensionality
  - Many relevant, redundant features
    - Correlated features (concurrence of words)
  - Document vectors are sparse

*(Gabrilovich & Markovitch04, Joachims98, Wang & Lochovsky04)*

# tf x idf (Salton&Buckley 1988)

- Assign a weight to each term *i* in each document *d*
-

$$w_{i,d} = tf_{i,d} \times \log(n / df_i)$$

$tf_{i,d}$ = frequency of term *i* in document *d*

$n$ = total number of documents

$df_i$ = the number of documents that contain term *i*

  - It increases with the number of occurrences *within d*
  - It increases with the rarity of *i across* the whole corpus

# Feature Selection Methods

$P(t_k, c_i)$ - the probability of term $t_k$ occurring of class $c_i$;

$N$ - the number of training documents

- Unsupervised
  - Document Frequency: $N*P(t_k)$
- Supervised
  - Information Gain
  - Chi-Squared
  - Odds Ratio
  - Bi-normal separation *(Forman 2003)*

# Measures (continued)

- Information Gain:
- Chi-Squared: $$N \cdot \frac{P(t_k,c_i)P(\overline{t_k},\overline{c_i}) - P(t_k,\overline{c_i})P(\overline{t_k},c_i)}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}$$
- Odds Ratio:

$$\frac{P(t_k|c_i)\cdot(1-P(t_k|\overline{c_i}))}{(1-P(t_k|c_i))\cdot P(t_k|\overline{c_i})}$$

- Bi-normal separation:

$$|F^{-1}(P(t_k|c_i)) - F^{-1}(P(t_k|\overline{c_i}))|$$

Where F is the cumulative probability function of the standard Normal distribution

# Empirical Studies

- One can significantly reduce the number of features (terms)
- Information Gain and Chi-Squared are most effective *(Yang and Pedersen97, Rogati and Yang02)*
- Odds ratio is reported to perform well when the data is skewed *(Mladenic and Grobelnik98)*
- Bi-normal separation outperforms IG and Chi-Squared in SVMs, especially when the class distribution is skewed *(Forman 2003)*

# Feature Interaction

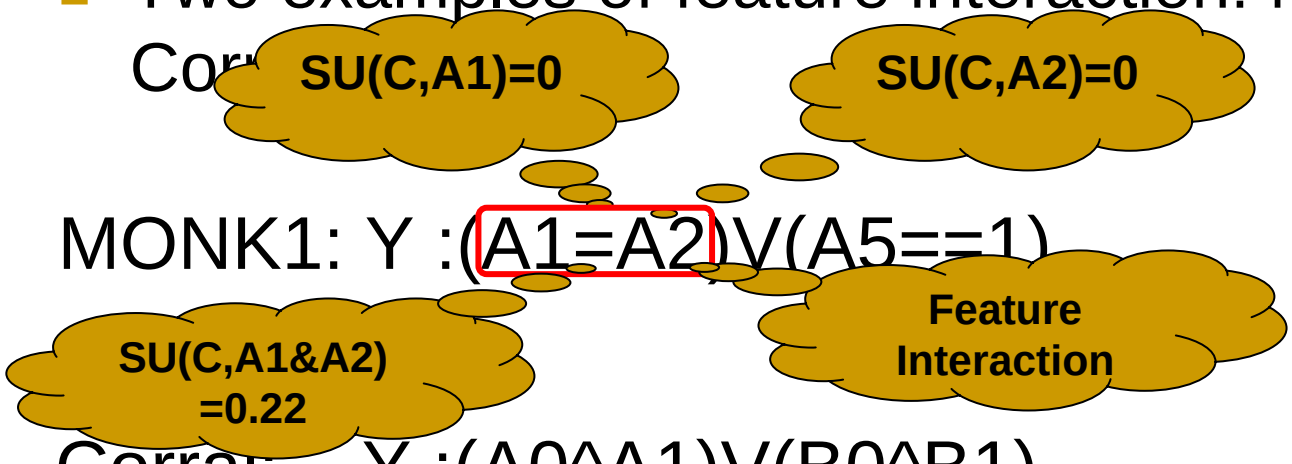
- Feature interaction is a phenomenon, in which features gain their relevance by interacting with other features.
- Definition:
  **F** = $F_1, F_2, \ldots, F_k$, let $\mathscr{C}$ denotes a metric measures the correlation of the class label with a feature or a feature set. Features $F_1, F_2, \ldots, F_k$ are said to interact with each other iff: for arbitrary partition $\mathscr{F} = \{\mathscr{F}_1, \mathscr{F}_2, \ldots, \mathscr{F}_l\}$ of **F**, where $l \geq 2$ and $\mathscr{F}_i \neq \emptyset$ for $\forall i \in [1, l]$ we have:

$$\mathscr{C}(\mathbf{F}) > \sum_{\mathscr{F}_i \in \mathscr{F}} \mathscr{C}(\mathscr{F}_i)$$

- An intrinsic character of feature interaction is its irreducibility. *(Jakulin & Bratko 2004)*

# Feature Interaction

- Two examples of feature interaction: MONK1 & Corral




Corral: Y :(A0^A1)V(B0^B1)

- Existing efficient feature selection algorithms cannot handle feature interaction very well

| | FCBF | CFS | ReliefF | FOCUS |
|---|---|---|---|---|
| Corral | $A_0, A_1, B_0, B_1, \mathbf{R}$ | $A_0, -, -, -, \mathbf{R}$ | $A_0, A_1, B_0, B_1, \mathbf{R}$ | $A_0, A_1, B_0, B_1$ |
| Monk1 | $-, -, A_5$ | $-, -, A_5$ | $A_1, A_2, A_5$ | $A_1, A_2, A_5$ |

# Feature Interaction

- Existing efficient feature selection algorithms usually assume feature independence.
- Others attempt to explicitly address Feature Interactions by finding them.
  - Finding out all feature interactions is impractical.
- Some existing efficient algorithm can only (partially) address low order Feature Interaction, 2 or 3-way Feature Interaction.

# Streaming Feature Selection

- Return to questions considered previously
  - Which features to consider, and how to judge them?
- Open versus closed domain
  - Is the collection of features a well-defined, finite list?

  Or
  - Is the collection of features “open” set in that choice of new features depends on success or failure of initial choices?
- Example
  - Text mining in which features are constructed from SQL queries of documents in database.
  - Next query depends on success of prior features.

# Infinite Feature Space

- Trend toward wider data sets

| Problem | Cases | Base Features |
|---|---|---|
| Credit default | 3,000,000 | 350 |
| Face recognition | 10,000 | 1,400 |
| Microarray | 1,000 | 10,000 |
| CiteSeer | 500 | 10,000,000 |

- Compounded by relevance of interactions

# Batch-Oriented Feature Selection

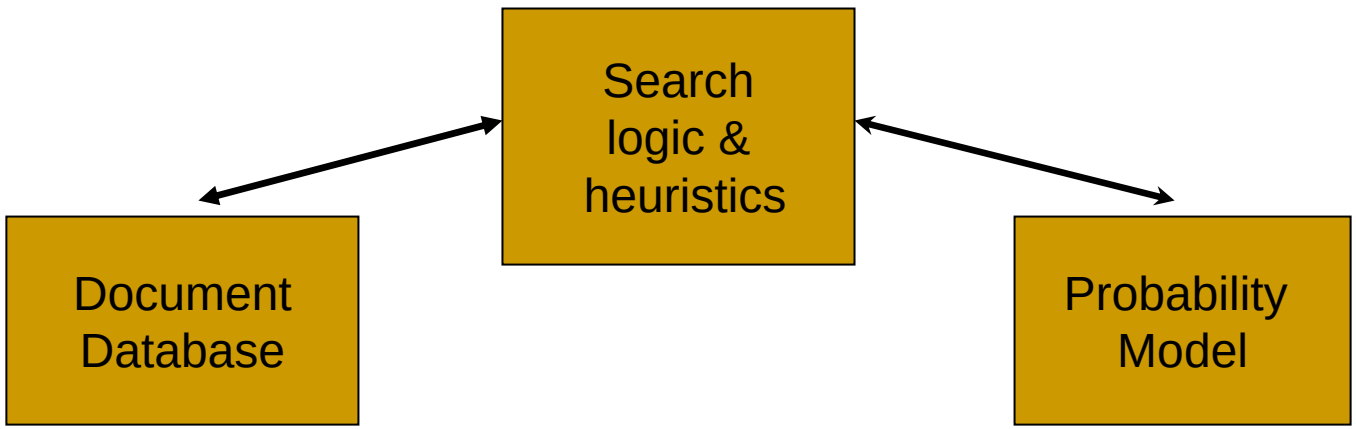
- Breadth-first selection is common
  - Greedy search identifies the single best feature from those available
- But it has issues…
  - Selection bias
    Search over many produces spurious “benefit”
  - Speed
    Slow, must consider every feature at each step
  - Scope
    Confinded to features defined in advance of search process and modeling

# Streaming Feature Selection

- Depth-first search
  - Accept any “adequate” feature that is discovered without “waiting for” the best
- Avoids problems
  - Selection bias no longer relevant because not picking the best in a batch comparison
  - Speed improves because only consider one feature at a time
  - Scope expands search space depending on successful choices
    - Role for feedback in directing search

# Code Factorization in Streaming Feature Selection

- Naturally factored problem
  - Postpone task of constructing feature until needed for evaluation
  - Separate task of recommending a feature to try from the task of evaluating a feature



# Expert: Feature Generation Algorithm

- Distinguish two flavors
- Context driven experts
  - If paper A cites paper B, then consider papers cited in paper B…
  - If a gene in this group is related to the response, then no need to consider genes in this category
- Parasitic "experts" piggyback on others
  - If some feature X is predictive, then consider…
  - Simple transformations $\phi(X)$
  - Interactions of the form X * Z
  - Deep search of high-order combinations

# Feature Evaluation

- How to evaluate the features offered by one (or more) experts?
  - Different approach from that used in batch context
- View as a sequence of
  - Traditional: Hypothesis tests
  - "Investment opportunities"
- Key concern?
  - Avoid features that are not informative out-of-sample (avoid over-fitting)
  - Miss features that would be informative out-of-sample (lack of power)
- Perspective
  - Avoid over-fitting while retaining as much power as can

# Classical Criterion

- Notation
  - Offered *fixed* collection of m features
  - R(m) counts chosen features (observable r.v.)

    $$R(m) = V_\beta(m) + S_\beta(m)$$
  - $V_\beta(m)$ counts incorrect picks, "false positives"
  - $S_\beta(m)$ counts correctly used "true positives"
  - Want large $S_\beta(m)$ and small $V_\beta(m)$
- Conservative approach
  - Control chance for *any* false positive,

    $$FWER = Pr(V_\beta(m) > 0) \leqslant \alpha$$
  - Eg: Bonferroni controls FWER but has low power

# False Discovery Rate

- Better batch method
- FDR (Benjamini&Hochberg)
  - Control *rate* of false positives rather than chance of any false positive

    $$FDR(m) = E(V_{\beta}(m)/R(m))$$
  - If picking lots of features, OK that some are false positives
- Procedure
  - B&H provide a step-down testing procedure
  - Procedure requires ordered collection of p-values, one for for every feature

    $$\text{p-value} = \Pr(\text{observed test stat} \mid H_0)$$

    with null hypothesis that each feature adds no value

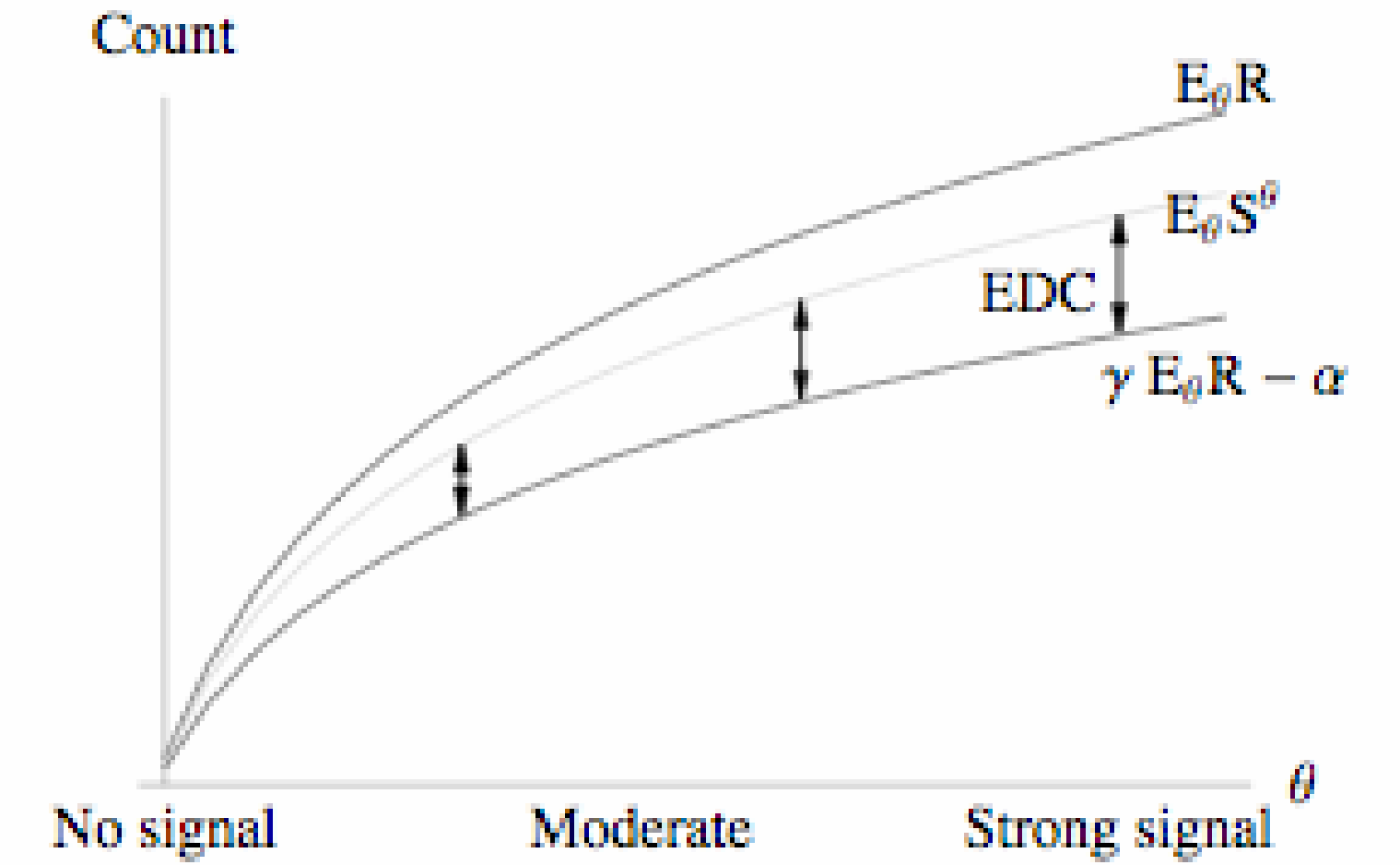
# Excess Discovery Count

- Streaming criterion
  - Accommodates stream of features rather than just a single batch of features
- EDC counts true features found above a fraction $\gamma$ of the total number chosen R(m)

$$EDC(m) = E[S_{\beta}(m) - \gamma R(m)] + \alpha$$

- A selection procedure "controls EDC" if EDC $\geqslant 0$
  - Typical choices set $\alpha = 0.05$ and $\gamma = 0.95$
- Comparison to FDR
  - Focus on correctly chosen features
  - Ratio of expected values rather than expected value of ratio
  - Avoids associated complications

# EDC provides guarantee



# Alpha-investing Procedures

- Feature selection procedure
  - Sequential testing procedure for evaluating a sequence of features
- Procedure "invests" in feature choices
  - Has wealth W(m) after considering m features
  - When offered a candidate feature, procedure determines how much of current wealth to invest
    - If the feature "seems good", invests heavily
    - If the feature "appears weak", does not invest
  - Methods that invest wisely earn wealth that in turn allows future investment

# Alpha-Investing

- Requires p-value that measures contribution of feature to model
  - Wealth measured on scale of p-values
- When considering the $m^{th}$ feature
  - Test null hypothesis

    $H_0$: This feature adds no value

    at level up to current wealth

    $\alpha_m \leqslant W(m-1)$
  - If p-value for this feature is small, $p_m \leqslant \alpha_m$, then reject $H_0$
    - Add $m^{th}$ feature to model
    - Increase wealth by payout $\omega - p_m$
  - If p-value $p_m > \alpha_m$ then do not reject $H_0$ and
    - Decrease wealth by $\alpha_m$

# Alpha Investing Controls EDC

- Theorem
  - Given suitable choices for the initial wealth and payoff $\omega$, alpha-investing controls EDC

$$E_M[\ \mathrm{EDC}(M) \geqslant 0\ ]\ ,$$

    over all stopping times M.
- Key requirement isolates p-values
  - Need for the evaluation tests of candidate features to be "honest" in sense that

$$\Pr(\text{incorrectly use feature}) \leqslant \alpha_m$$

# Alpha-investing is flexible

- Taylor the evaluation of features to the "reputation" of the expert
- Invest more heavily in the recommendations of experts that know more or experts whose choices have resulted in valuable improvements in the model
- Invest conservatively in other experts, at least until they have a reputation

# Strategies: No Domain Knowledge

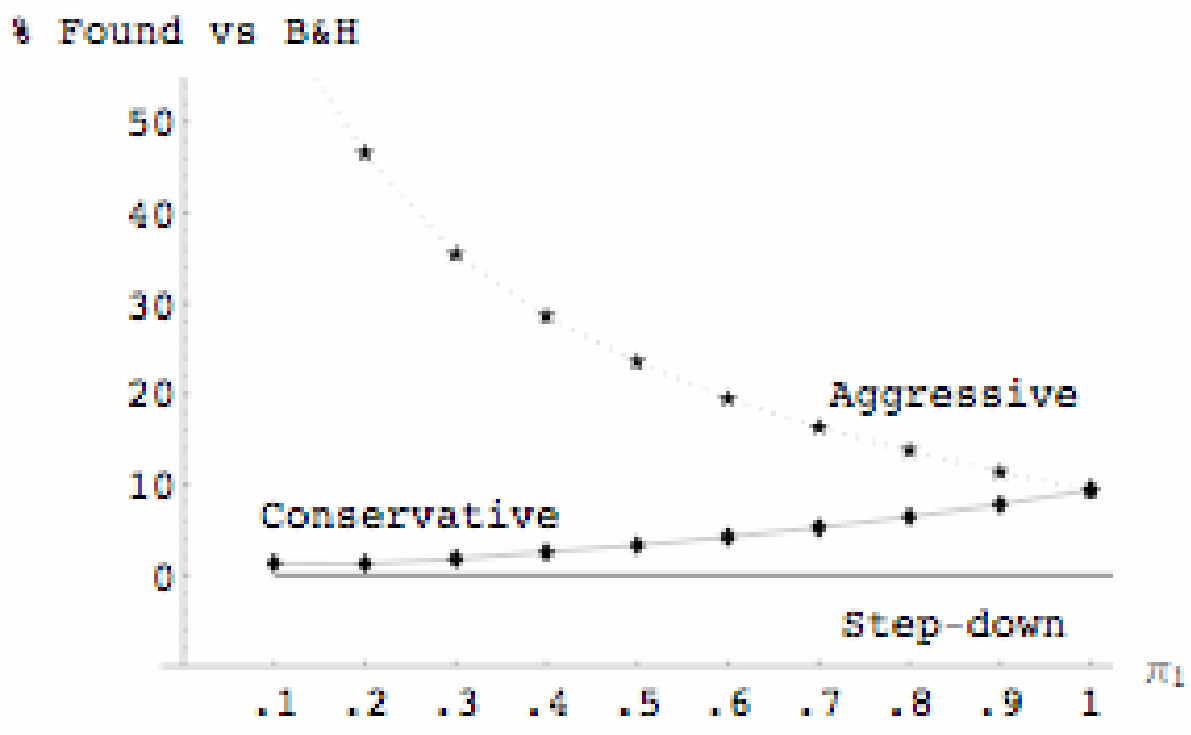
- If know little of the domain, then alpha-investing can operate "conservatively"
  - Expert offers little insight, poor reputation
- Invest small fraction of wealth on each feature to conserve for future opportunities
- In batch context, conservative alpha-investing replicates the FDR procedure
  - Test all m features at level $\alpha/m$
  - Test remaining features at level 2 $\alpha/m$
  - Continue up list of features

# Strategies: With Domain Knowledge

- Aggressive investing
  - Genetics expert recommends attributes
  - Credit modeler experience in default
- Assume knowledge of feature space places an ordering on the candidate features such that most interesting tested first
- Invest heavily in leading features since these are most likely to be useful
  - Test first feature at level $\alpha/2$
  - If not used, test second at level $\alpha/4$
  - If not used, test third at level $\alpha/8$ …

# Example

- Fraction $\pi_1$ predictors hidden in 200 features
- Sequential search using 2 investing schemes

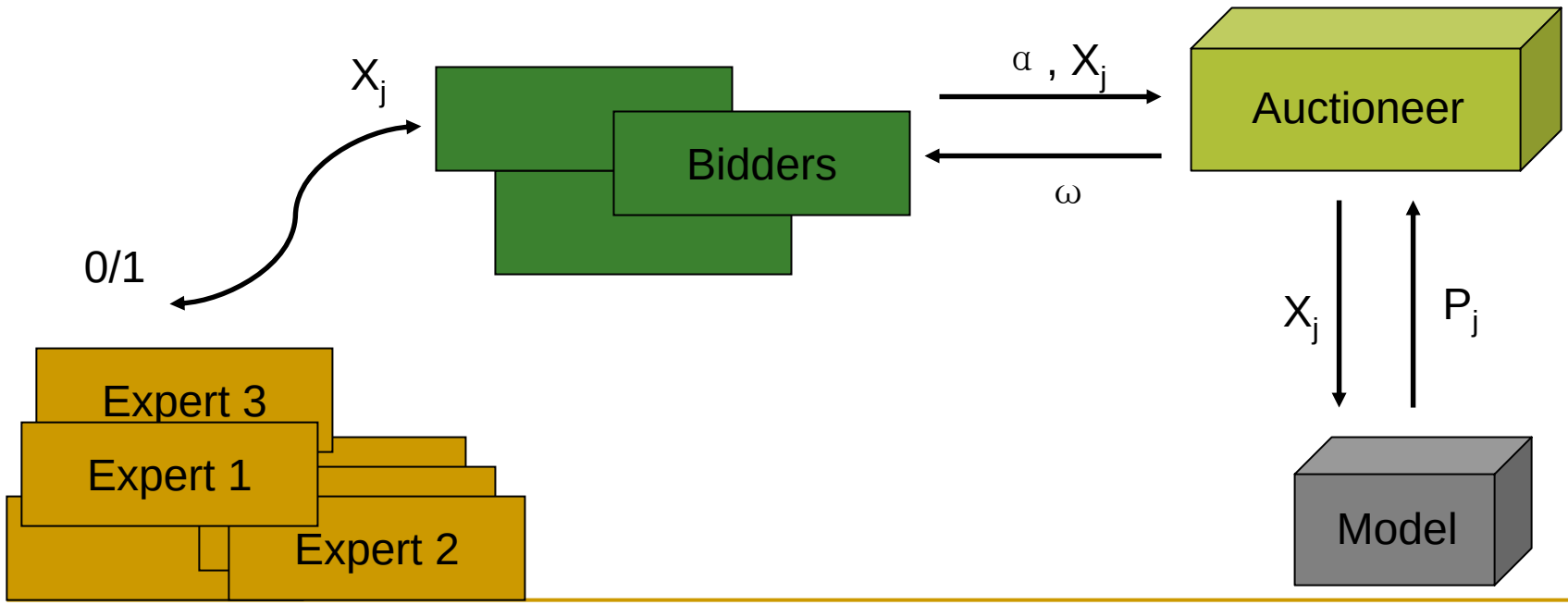


# Combining Experts in Auction

- Predictor auction matches up
  - Experts: Feature generating strategies
  - Bidders: Alpha-investing rules
- Governed by auctioneer
  - Bidders wager on features offered by collection of experts
  - Auctioneer “accepts” feature with largest bid value
- Evaluates feature
  - Tests this feature in current model
  - If feature is accepted, rewards those bidders
  - If feature is declined, bidders lose investment

# Feature Auction Schematic

- Factor problem
  - Domain knowledge in the experts
  - Investing strategies in bidders
  - Feature assessment in model



# Streaming Feature Selection

- New approach to old problem
- Experts offer stream of features
  - Domain specific, exploiting substantive structure
  - Parasitic, exploiting choices of others
- Alpha-investing rules allow bidders to evaluate the features in the stream
  - Allows infinite stream of features
  - Taylor investing strategy to problem
- Auctions combine multiple experts with variety of bidding strategies

# Acknowledgements


- Huan Liu's part is joint work with Lei Yu, Lance Parsons, Zheng Zhao, and Lei Tang

# References

## Section One


Almuallim, H., and Dietterich, T. G. 1994. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence* 69(1-2):279–305.

Cant-Paz, E.; Newsam, S.; and Kamath, C. 2004. Feature selection in scientific applications. In *Proceedings of the Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 788–793.

Dash, M., and Liu, H. 1997. Feature selection for classification. *Intelligent Data Analysis: An International Journal* 1(3):131–156.

Dash, M., and Liu, H. 2000. Feature selection for clustering. In *Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, (PAKDD-2000). Kyoto, Japan*, 110–121. Springer-Verlag.

Dash, M., and Liu, H. 2003. Consistency-based search in feature selection. *Artificial Intelligence* 151(1-2):155–176.

Devaney, M., and Ram, A. 1997. Efficient feature selection in conceptual clustering. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 92–97.

Ding, C., and Peng, H. 2003. Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the Computational Systems Bioinformatics conference (CSB'03)*, 523–529.

Dy, J. G., and Brodley, C. E. 2000. Feature subset selection and order identification for unsupervised learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 247–254.

Dy, J. G., and Brodley, C. E. 2004. Feature selection for unsupervised learning. *J. Mach. Learn. Res.* 5:845–889.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3:1157–1182.

Hall, M. A. 2000. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 359–366.

Jebara, T., and Jaakkola, T. 2000. Feature selection and dualities in maximum entropy discrimination. In *In Uncertainity In Artificial Intellegence*.

John, G. H.; Kohavi, R.; and Pfleger, K. 1994. Irrelevant feature and the subset selection problem. In *Proceedings of the Eleventh International Conference on Machine Learning*, 121–129.

Kira, K., and Rendell, L. 1992. A practical approach to feature selection. In Sleeman, and Edwards, P., eds., *Proceedings of the Ninth International Conference on Machine Learning (ICML-92)*, 249–256. Morgan Kaufmann.

Kohavi, R., and John, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97(1-2):273–324.

Koller, D., and Sahami, M. 1996. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 284–292.

Li, T.; Zhang, C.; and Ogihara, M. 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20:2429–2437.

Liu, H., and Motoda, H. 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers.

Liu, H., and Setiono, R. 1996. A probabilistic approach to feature selection - a filter solution. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 319–327.

Liu, H., and Setiono, R. 1997. Feature selection via discretization. *IEEE Trans on Knoweldge and Data Engineering* 9(4):642–645.

Liu, H., and Yu, L. 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering* 17:491–502.

Liu, H.; Motoda, H.; and Yu, L. 2004. A selective sampling approach to active feature selection. *Artificial Intelligence* 159:49–74.

Miller, A. 2002. *Subset Selection in Regression*. Chapman & Hall/CRC, 2 edition.

Mitra, P.; Murthy, C. A.; and Pal, S. K. 2002. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(3):301–312.

Modrzejewski, M. 1993. Feature selection using rough sets theory. In Brazdil, P., ed., *Proceedings of the European Conference on Machine Learning*, 213–226.

Pudil, P., and Novovicová, J. 1998. Novel methods for subset selection with respect to problem knowledge. *IEEE Intelligent Systems* 13(2):66–74.

Robnik-Sikonja, M., and Kononenko, I. 2003. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning* 53(1-2):23–69.

Xing, E.; Jordan, M.; and Karp, R. 2001. Feature selection for high-dimensional genomic microarray data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 601–608.

Xiong, M.; Fang, Z.; and Zhao, J. 2001. Biomarker identification by feature wrappers. *Genome Research* 11:1878–1887.

Yu, L., and Liu, H. 2004a. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* 5:1205–1224.

Yu, L., and Liu, H. 2004b. Redundancy based feature selection for microarray data. In *Proceedings of the Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 737–742.

## Section Two

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000) Adapting to unknown sparsity by controlling the false discovery rate. *Tech. Rep. 2000–19*, Dept. of Statistics, Stanford University, Stanford, CA.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csàki), 261–281. Budapest: Akad. Kiàdo.

Barron, A. R., Rissanen, J. and Yu, B. (1998) The minimum description length principle in coding and modeling. *IEEE Trans. on Info. Theory*, **44**, 2743–2760.

Donoho, D. (2002) Kolmogorov complexity. *Tech. rep.*, Stanford University, Stanford, CA.

Donoho, D. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *Journal of the Amer. Statist. Assoc.*, **90**, 1200–1224.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Foster, D. P. and George, E. I. (1994) The risk inflation criterion for multiple regression. *Annals of Statistics*, **22**, 1947–1975.

Foster, D. P., Stine, R. A. and Wyner, A. J. (2002) Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Trans. on Info. Theory*, **48**, 1713–1720.

George, E. I. and Foster, D. P. (2000) Calibration and empirical bayes variable selection. *Biometrika*, **87**, 731–747.

Grünwald, P., Myung, I. J. and Pitt, M. A. (eds.) (2005) *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Hansen, M. H. and Yu, B. (2001) Model selection and the principle of minimum description length. *Journal of the Amer. Statist. Assoc.*, **96**, 746–774.

Johnstone, I. M. and Silverman, B. W. (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, **32**, 1594–1649.

Mallows, C. L. (1973) Some comments on $C_p$. *Technometrics*, **15**, 661–675.

Miller, A. J. (2002) *Subset Selection in Regression (Second Edition)*. London: Chapman& Hall.

Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416–431.

Rissanen, J. (1986) Stochastic complexity and modeling. *Annals of Statistics*, **14**, 1080–1100.

Shen, X. and Ye, J. (2002) Adaptive model selection. *Journal of the Amer. Statist. Assoc.*, **97**, 210–221.

Tibshirani, R. and Knight, K. (1999) The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statist. Soc., Ser. B*, **61**, 529–546.

Ye, J. (1998) On measuring and correcting for the effects of data mining and model selection. *Journal of the Amer. Statist. Assoc.*, **93**, 120–131.

Zhang, C.-H. (1997) Empirical Bayes and compound estimation of normal means. *Statistica Sinica*, **7**, 181–193.

Cun-Hui Zhang (2001) General empirical Bayes wavelet methods and exactly adaptive minimax estimation. Unpublished working paper.

## Section Three

Dy, J. G.; Brodley, C. E.; Kak, A. C.; Broderick, L. S.; and Aisen, A. M. 2003. Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(3):373–378.

Foreman, G. 2004. A pitfall and solution in multi-class feature selection for text classification. In *Proceedings of the Twenty-First International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3:1289–1305.

Gabrilovich, E., and Markovitch, S. 2004. Text categorization with many redundant features: Using aggressive feature selection to make svms competitive with c4.5. In *Proceedings of The Twenty-First International Conference on Machine Learning*, 321–328. Banff, Alberta, Canada: Morgan Kaufmann.

Gilad-Bachrach, R.; Navot, A.; and Tishby, N. 2004. Margin based feature selection - theory and algorithms. In *ICML '04: Twenty-first international conference on Machine learning*. ACM Press.

Jakulin, A., and Bratko, I. 2004. Testing the significance of attribute interactions. In *ICML '04: Twenty-first international conference on Machine learning*. ACM Press.

Joachims, T. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, 137–142. Springer-Verlag.

Liu, H., and Motoda, H., eds. 2001. *Instance Selection and Construction for Data Mining*. Boston: Kluwer Academic Publishers.

Mladenic, D., and Grobelnik, M. 1999. Feature selection for unbalanced class distribution and Naive Bayes. In *Proceedings of Sixteenth International Conference on Machine Learning*, 258–267.

Reunanen, J. 2003. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research* 3:1371–1382.

Robnik-Sikonja, M., and Kononenko, I. 1999. Attribute dependencies, understandability and split selection in tree based models. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, 344–353. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rogati, M., and Yang, Y. 2002. High-performing feature selection for text classification. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, 659–661. New York, NY, USA: ACM Press.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24(5):513–523.

Wang, G., and Lochovsky, F. H. 2004. Feature selection with conditional mutual information maximin in text categorization. In *CIKM '04: Proceedings of the Thirteenth ACM conference on Information and knowledge management*, 342–349. New York, NY, USA: ACM Press.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 412–420. Morgan Kaufmann Publishers Inc.

Yu, L., and Liu, H. 2004. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* 5:1205–1224.

## Section Four

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statist. Soc., Ser. B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Braun, H. I. (ed.) (1994) *The Collected Works of John W. Tukey: Multiple Comparisons*, vol. VIII. New York: Chapman & Hall.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Efron, B. (2001) Selection criteria for scatterplot smoothing. *Annals of Statistics*, **29**, 470–504.

Gupta, M. and Ibrahim, J. G. (2005) Towards a complete picture of gene regulation: using Bayesian approaches to integrate genomic sequence and expression data. *Tech. rep.*, University of North Carolina, Chapel Hill, NC.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Meinshausen, N. and Buehlmann, P. (2004) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence. *Tech. Rep. 121*, ETH Zurich, http://stat.ethz.ch/ nicolai/.

Meinshausen, N. and Rice, J. (2004) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. To appear, *Annals of Statistics*.

Simes, R. J. (1986) An improved bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Foster, D. P. and Stine, R. A. (1996) Variable selection via information theory. *Tech. Rep. Discussion Paper 1180*, Center for Mathematical Studies in Economics and Management Science, Northwestern University, Chicago.