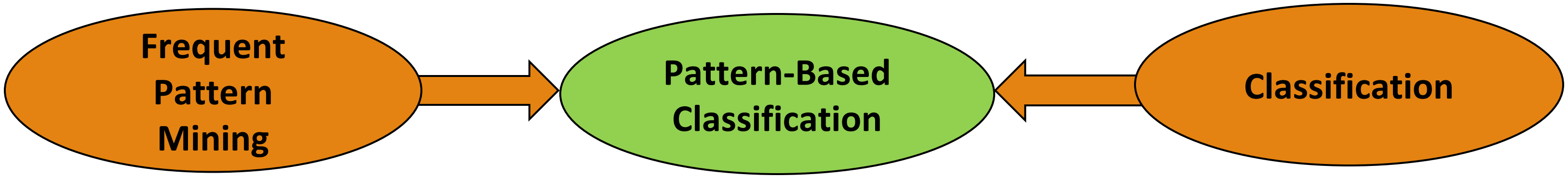
CS 412 Office Hour

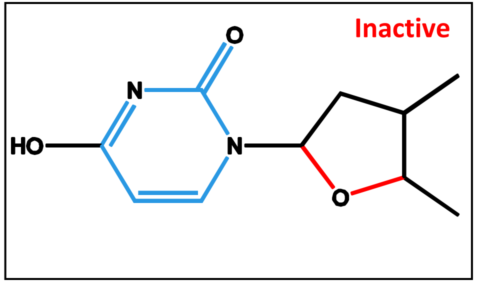
Apr 24, 2019

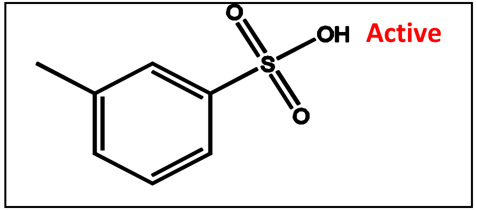
Why Pattern-Based Classification?

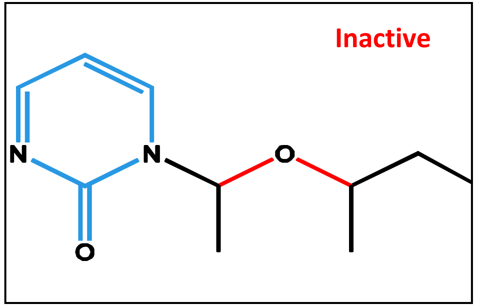


- **The Role of Patterns in Classification Models**
  - Rule based classifier using patterns (Associative Classification)
  - Feature construction
    - Higher order, often has more discriminative power e.g., single word → phrase (apple pie, Apple iPad)
    - Complex data modeling such as graphs, sequences

Pattern-Based Classification on Graphs



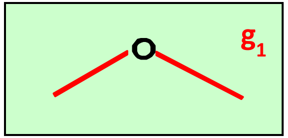


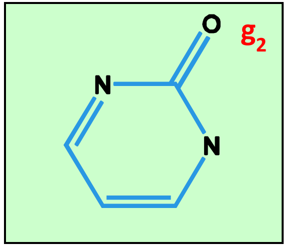


Mining

min_sup=2

Frequent subgraphs





Transform

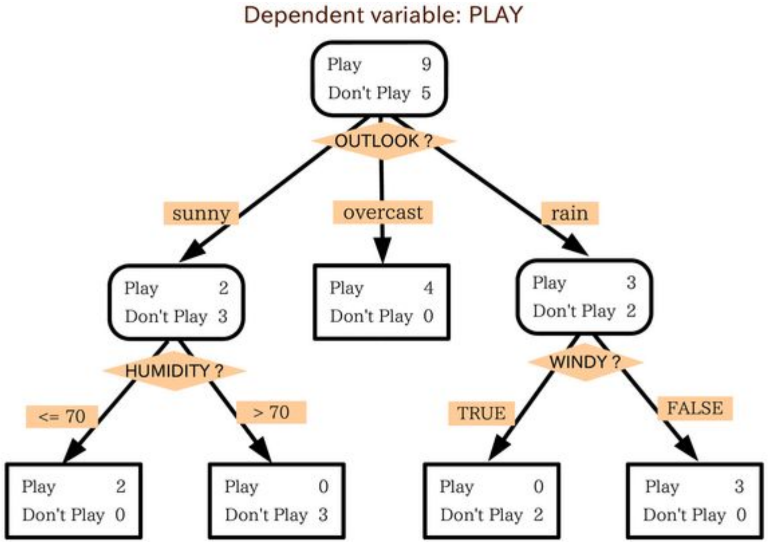
Use frequent patterns as features for classification

g_1	g_2	Class
1	1	0
0	0	1
1	1	0

Associative Classification: CBA and CMAR

Classification Rule Mining

- Association Rule Mining
  - Mining rules between itemsets
  - Every rule is associated with frequency and confidence
- Classification Rule Mining
  - Mining rules between attributes and the class label
  - A path from root to leaf on a decision tree can be seen as a classification rule



Rule 1: Outlook=rain $\wedge$ Windy =False $\Rightarrow$ Play=True
Rule 2: Outlook=rain $\wedge$ Windy =True $\Rightarrow$ Play=False
Rule 3: Outlook=overcast $\Rightarrow$ Play=True
Rule 4: Outlook=sunny $\wedge$ Humidity <=70 $\Rightarrow$ Play=True
Rule 5: Outlook=sunny $\wedge$ Humidity >70 $\Rightarrow$ Play=False

A decision tree defines a constrained set of rules: attributes that appear on parent nodes must appear before child nodes, e.g. rule “Windy=False $\wedge$ Humidity <=70 $\Rightarrow$ Play=True” will conflict with this set of rules.

CBA: Classification Based on Associations

- Mine high-confidence, high-support class association rules
  - LHS: Conjunctions of attribute-value pairs; RHS: Class labels
    $p_1 \wedge p_2 \ldots \wedge p_l \rightarrow$ "$A_{class\text{-}label} = C$" (confidence, support)
- Rank rules in descending order of confidence and support
  - Classification: Apply the first rule that matches a test case; otherwise, apply the default rule
- Effectiveness:More accurate than single decision tree
  - Why? – Exploring high confident associations among multiple attributes may overcome some constraints introduced by considering one attribute at a time
  - Drawback: the number of possible association rules is often large

CMAR: Classification Based on Multiple Association Rules

- Rule pruning whenever a rule is inserted
  - Given two rules R_1 and R_2, if the antecedent of R_1 is more general than that of R_2 and $\mathrm{conf}(R_1) \geq \mathrm{conf}(R_2)$, then prune R_2
  - Prunes rules for which the rule antecedent and class label are not positively correlated based on the χ^2 test of statistical significance
- Classification based on generated/pruned rules
  - If only *one rule* satisfies tuple X, assign the class label of the rule
  - If a *rule set* S satisfies X
    - Divide S into groups according to class labels
    - Use a weighted χ^2 measure to find the strongest group of rules based on the statistical correlation of rules within a group
    - Assign X the class label of the strongest group
- CMAR improves model construction efficiency and classification accuracy

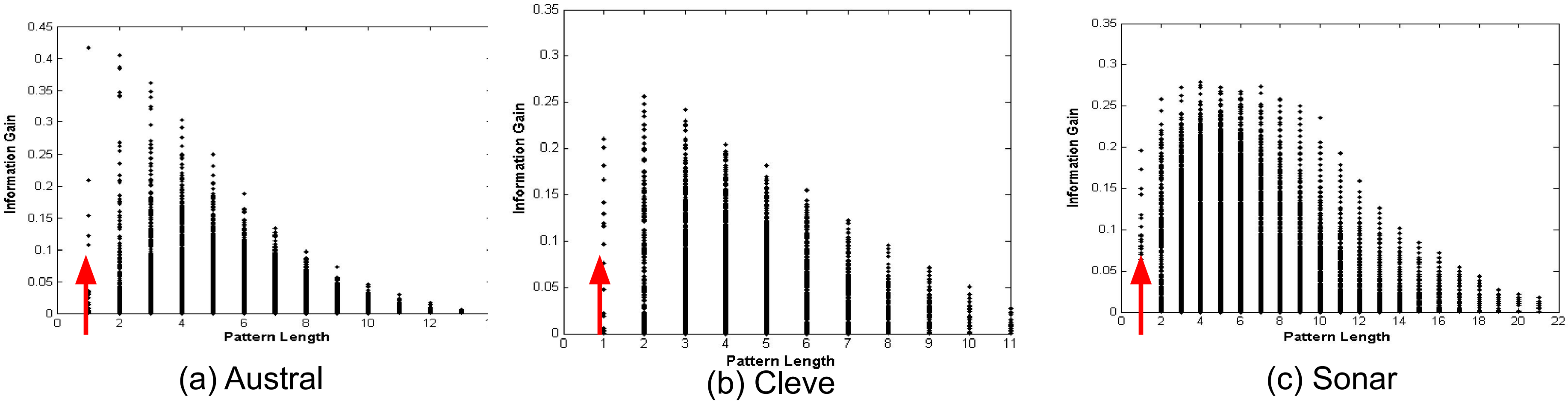
Discriminative Pattern-Based Classification

Discriminative Pattern-Based Classification

- Discriminative patterns as features for classification (Cheng et al., ICDE'07)
- **Principle:** Mining discriminative frequent patterns as high-quality features and then apply any classifier
- **Framework (PatClass)**
  - Feature construction by *frequent itemset mining*
  - Feature selection (e.g., using Maximal Marginal Relevance (MMR))
    - Select discriminative features (i.e., that are relevant but minimally similar to the previously selected ones)
    - Remove redundant or closely correlated features
  - Model learning
    - Apply a general classifier, such as SVM or C4.5, to build a classification model

On the Power of Discriminative Patterns

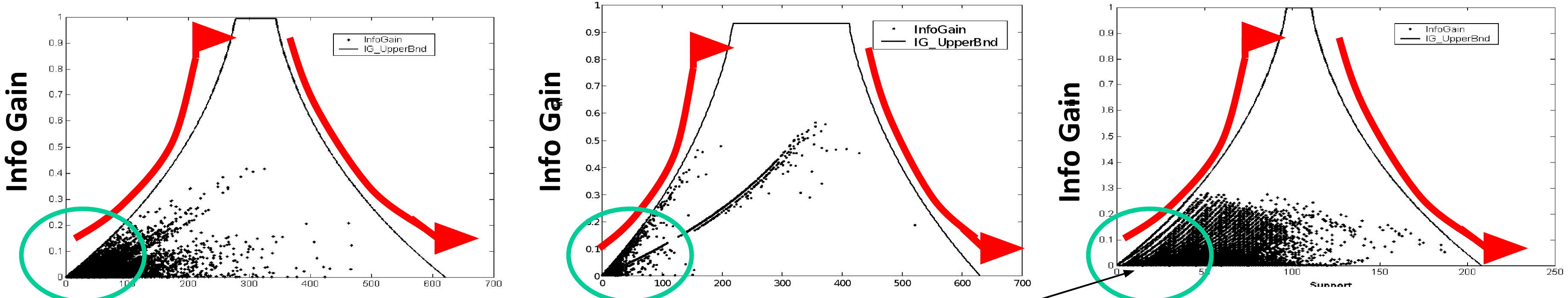
- K-itemsets are often more informative than single features (1-itemsets) in classification
- Computation on real datasets shows: The discriminative power of k-itemsets (for k > 1 but often ≤ 10) is higher than that of single features



Information Gain vs. Pattern Length

Information Gain vs. Pattern Frequency

- Computation on real datasets shows: Pattern frequency (but not too frequent) is strongly tied with the discriminative power (information gain)
- Information gain upper bound monotonically increases with pattern frequency



Frequency
(a) Austral

Frequency
(b) Breast

Low support, low info gain

Frequency
(c) Sonar

Information Gain Formula: $IG(C \mid X) = H(C) - H(C \mid X)$

Conditional entropy of study focus

Entropy of given data

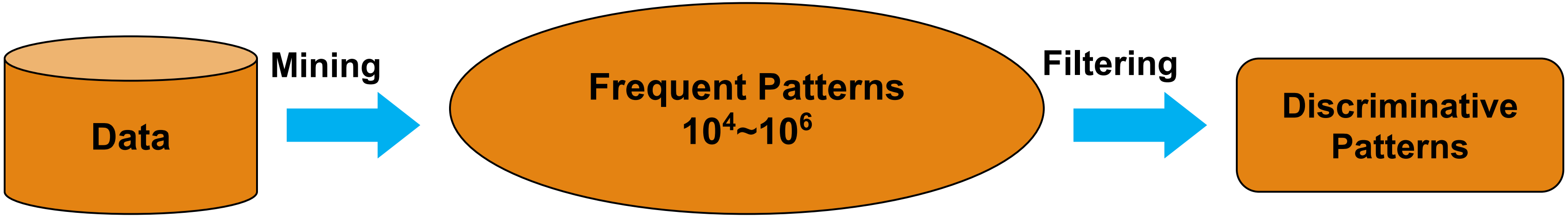
$$H(C) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$H(C \mid X) = \sum_j P(X = x_j) H(Y \mid X = x_j)$$

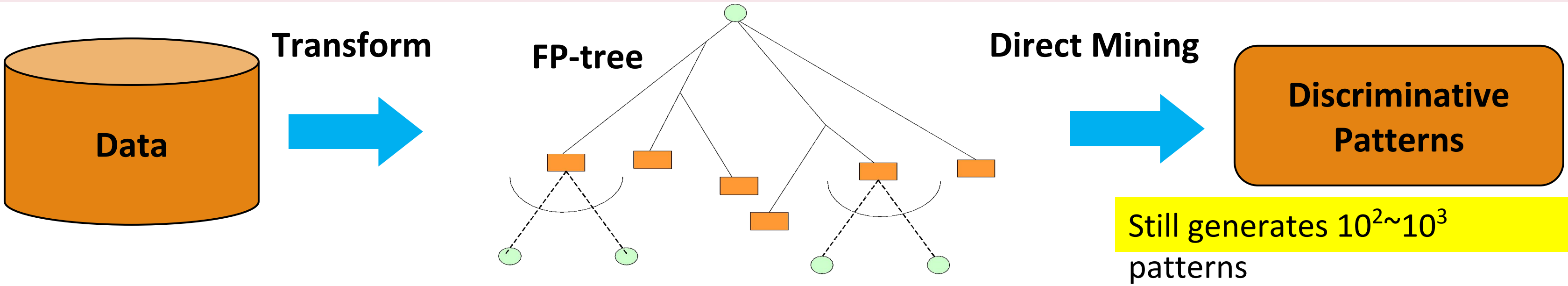
DPClass: Mining Concise Set of Discriminative Patterns for Classification

Mining Concise Set of Discriminative Patterns

Frequent pattern mining, then getting discriminative patterns: Expensive, large model



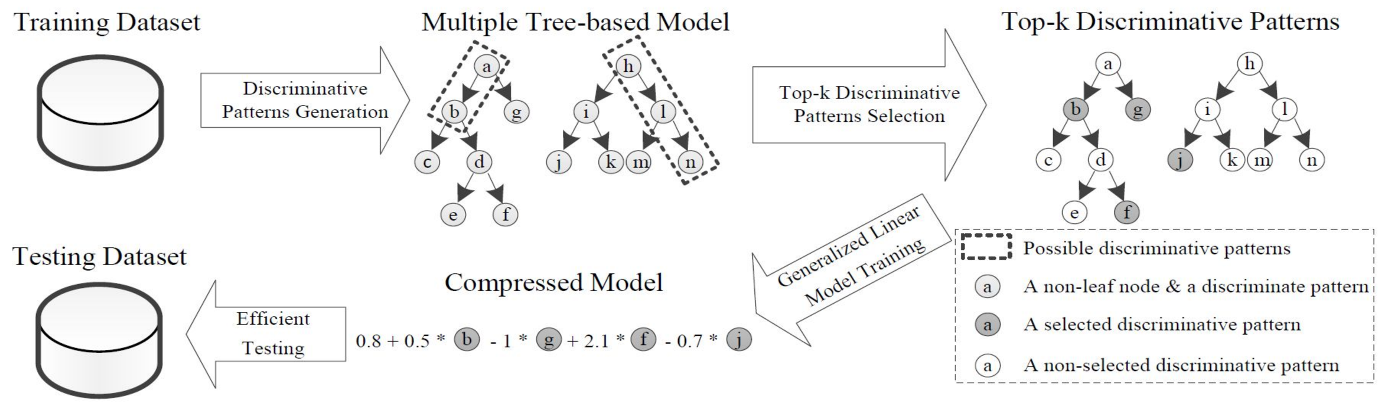
DDPMine (Cheng et al., ICDE'08): Direct mining of discriminative patterns: Efficient



Still generates 10^2~10^3 patterns

DPClass (Shang et al., SDM'16): A better solution (see the next page)—efficient, effective & generating a very limited number of (such as only 20 or so) patterns

DPClass: Discriminative Pattern-Based Classification



Input: A feature table for training data

- Adopt every prefix path in an (extremely) random forest as a candidate pattern
  - The split points of continuous variables are automatically chosen by random forest → No discretization!
- Run top-k (e.g., top-20) pattern selection based on training data
- Train a generalized linear model (e.g., logistic regression) based on "bag-of-patterns" representations of training data

Explanatory Discriminative Patterns: Generation

- Example: For each patient, we have several uniformly sampled features as follows
  - Age (A): Positive integers no more than 60
  - Gender (G): Male or female
  - Lab Test 1 (LT1): Categorical values from (A, B, O, AB)
  - Lab Test 2 (LT2): Continuous values in [0..1]
- The positive label of the hypothetical disease will be given when at least one of the following rules holds
  - (age > 18) and (gender = Male) and (LT1 = AB) and (LT2 ≥ 0.6)
  - (age > 18) and (gender = Female) and (LT1 = O) and (LT2 ≥ 0.5)
  - (age ≤ 18) and (LT2 ≥ 0.9)
- Training: 10^5 random patients + add 0.1% noise
  - Flip the binary labels with 0.1% probability
- Testing: 5×10^4 random patients in test

Explanatory Discriminative Patterns: Evaluation

- Accuracy:
  - DPClass 99.99% (perfect)
  - DDPMine 95.64% (reasonable)
- Top-3 Discriminative Patterns:
  - DPClass generates a high quality model here:
    - (age > 18) and (gender = Female) and (LT1 = O) and (LT2 ≥ 0.496)
    - (age ≤ 18) and (LT2 ≥ 0.900)
    - (age > 18) and (gender = Male) and (LT1 = AB) and (LT2 ≥ 0.601)
  - DDPMine generates a relatively poor quality model here:
    - (LT2 > 0.8)
    - (gender = Male) and (LT1 = AB) and (LT2 ≥ 0.6) and (LT2 < 0.8)
    - (gender = Female) and (LT1 = O) and (LT2 ≥ 0.6) and (LT2 < 0.8)