Advanced, Pattern-Based Classification

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

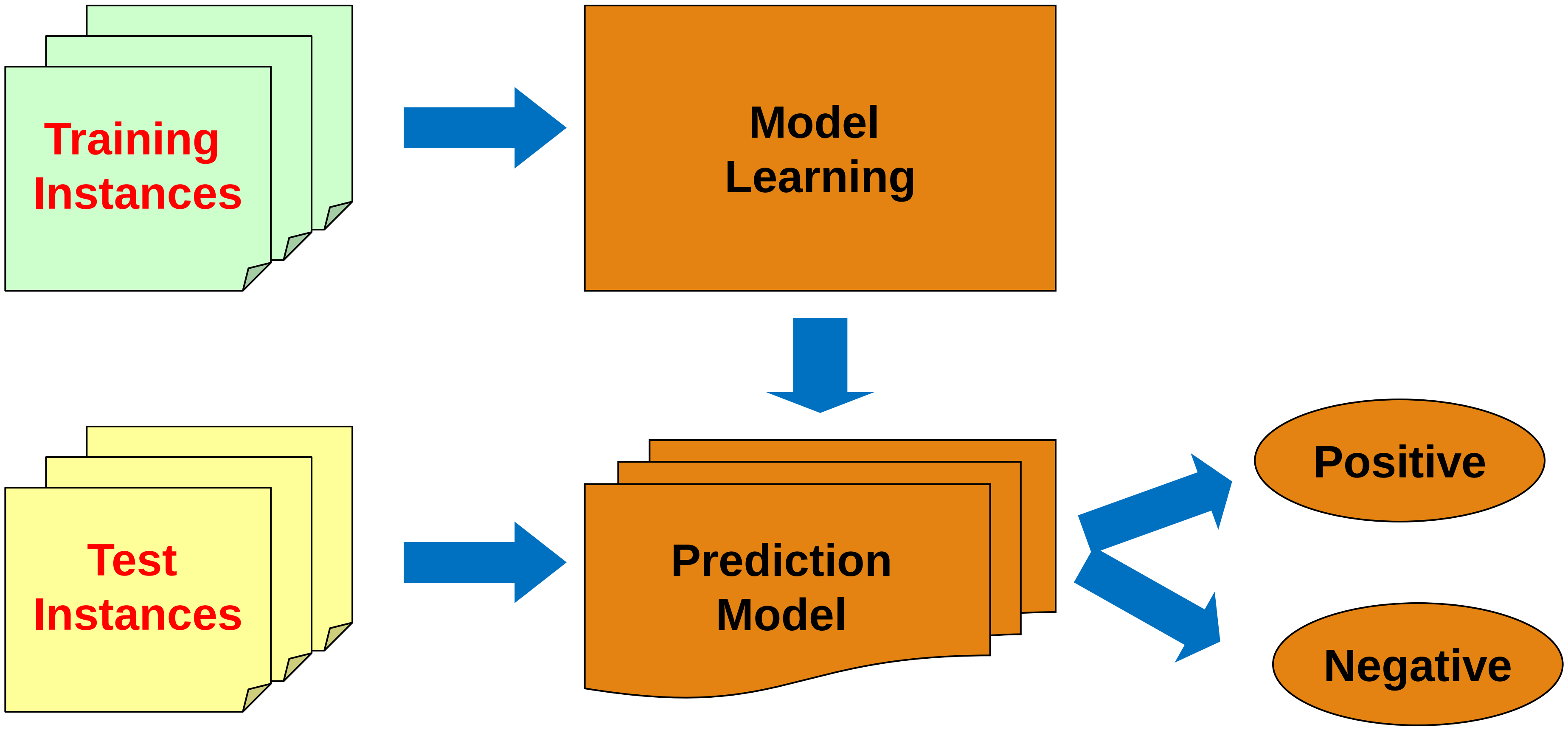
APRIL 20, 2017

Advanced Pattern-Based Classification

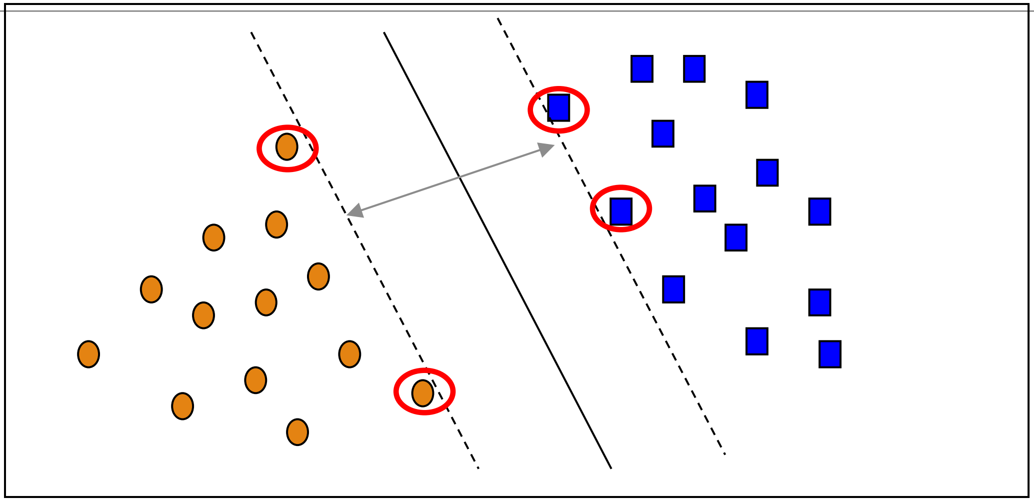
- Classification: Basic Concepts
- Pattern-Based Classification
- Associative Classification
- Discriminative Pattern-Based Classification
- Direct Mining of Discriminative Patterns
- DPClass: Effective but Concise Discriminative Patterns-Based Classification

Thanks to Hong Cheng@CUHK and Jingbo Shang @UIUC for their contributions

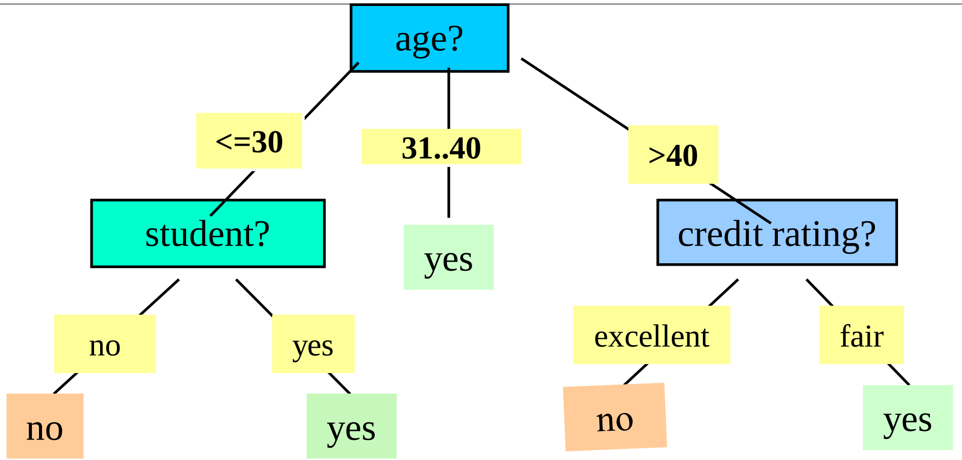
What Is Classification?



Typical Classification Methods

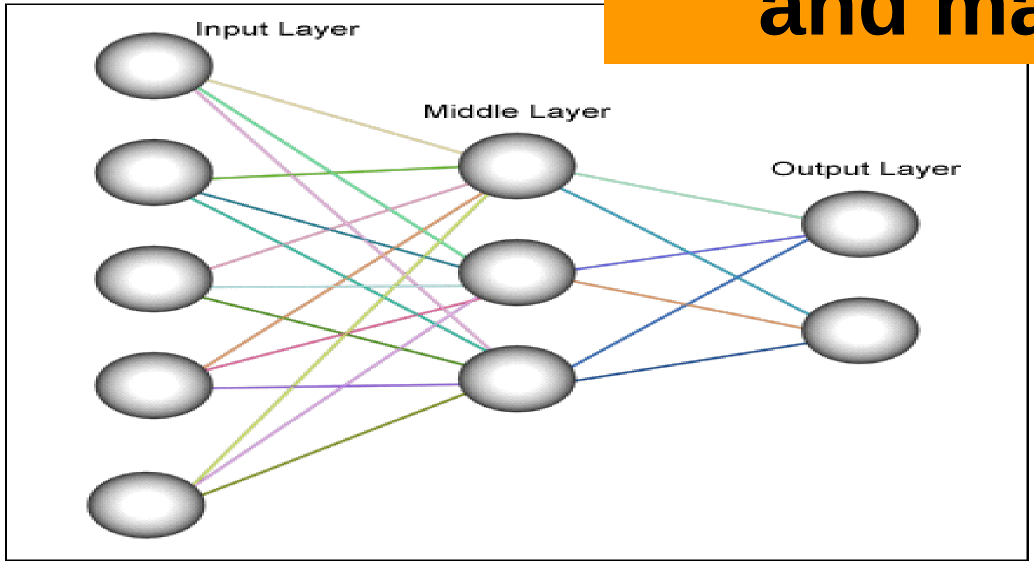


Support Vector Machi

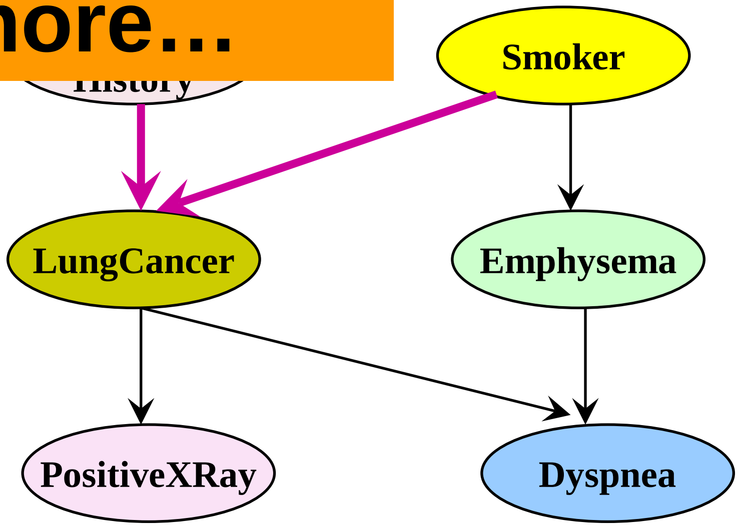


Decision Tree

and many more…

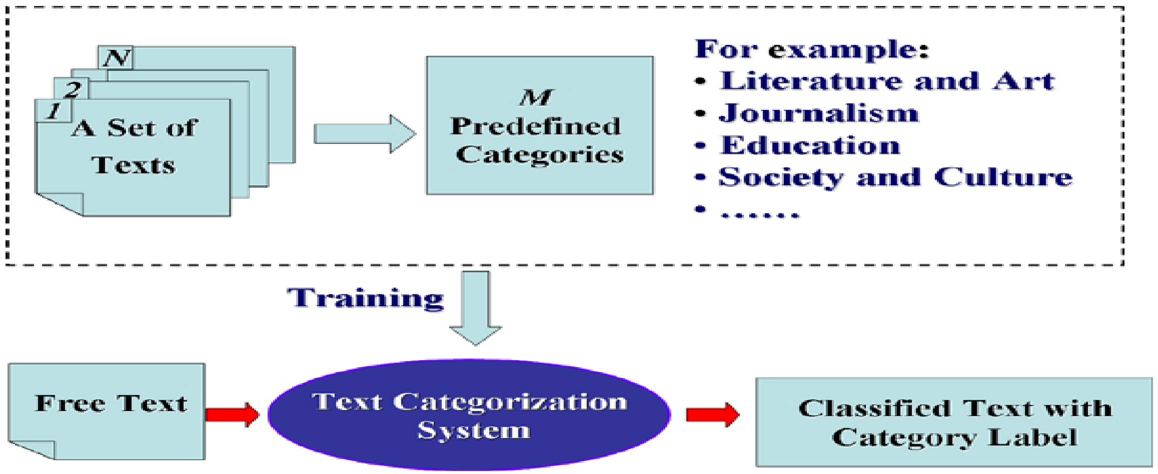


Neural Network

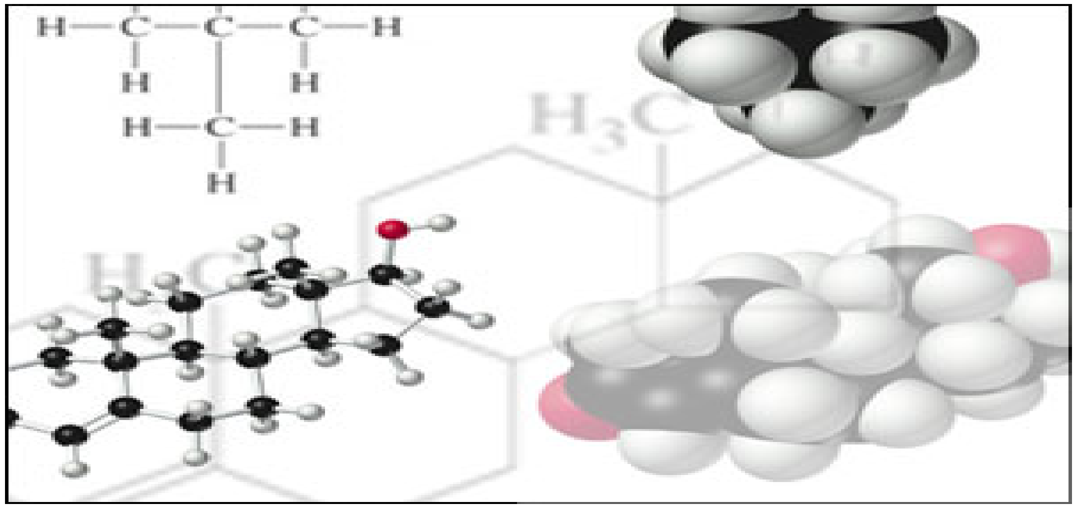


Bayesian Network

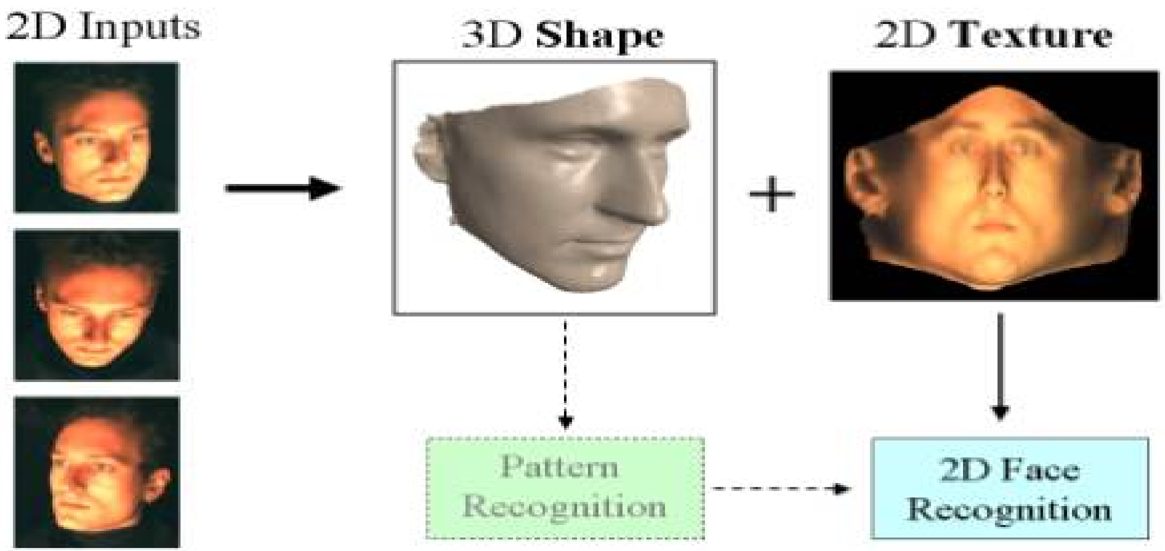
Numerous Classification Applications



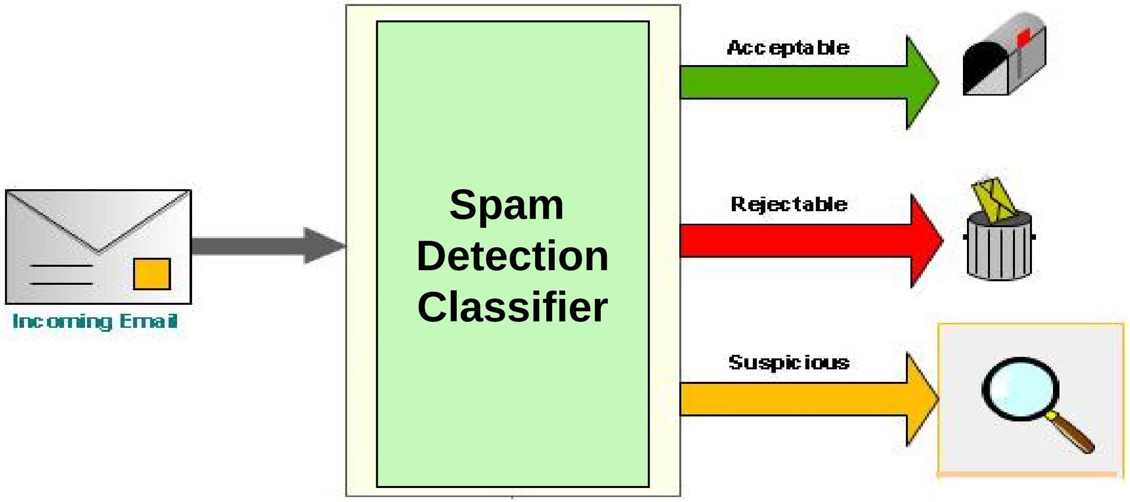
Text Categorization



Drug Design



Face Recognition



Spam Detection

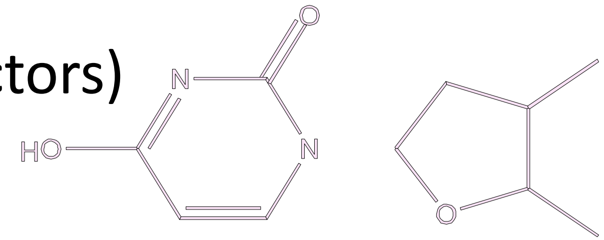
Advanced Pattern-Based Classification

- Classification: Basic Concepts
- Pattern-Based Classification
- Associative Classification
- Discriminative Pattern-Based Classification
- Direct Mining of Discriminative Patterns
- DPClass: Effective but Concise Discriminative Patterns-Based Classification

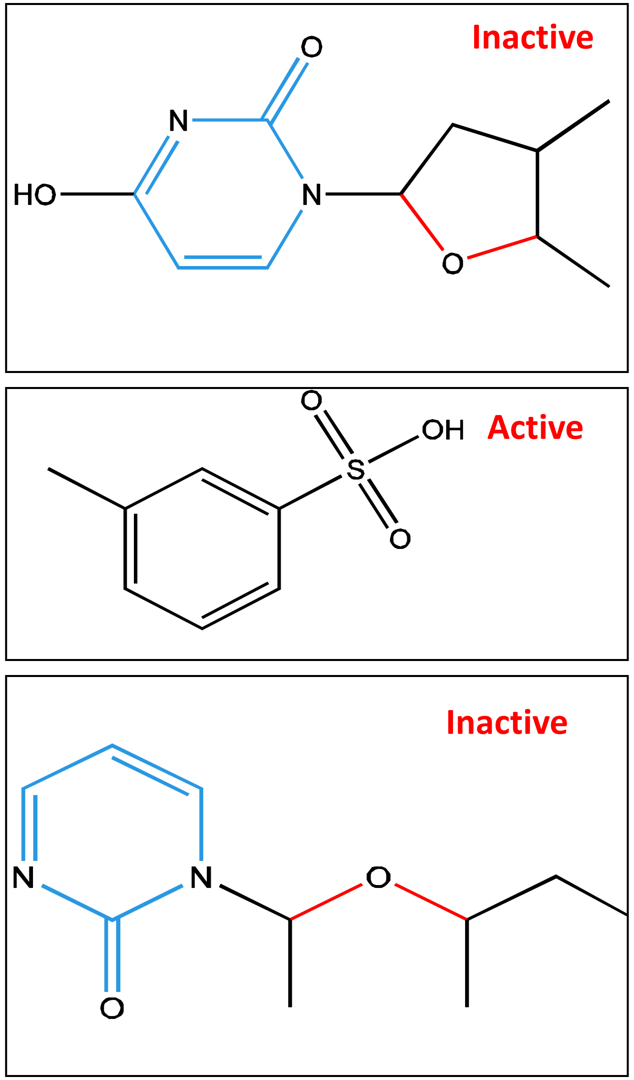
Thanks to Hong Cheng@CUHK and Jingbo Shang @UIUC for their contributions

Pattern-Based Classification, Why?

- **Pattern-based classification:** An integration of both themes
- **Why pattern-based classification?**
  - **Feature construction**
    - Higher order; compact; discriminative
    - E.g., single word → phrase (Apple pie, Apple i-pad)
  - **Complex data modeling**
    - Graphs (no predefined feature vectors)
    - Sequences
    - Semi-structured/unstructured Data



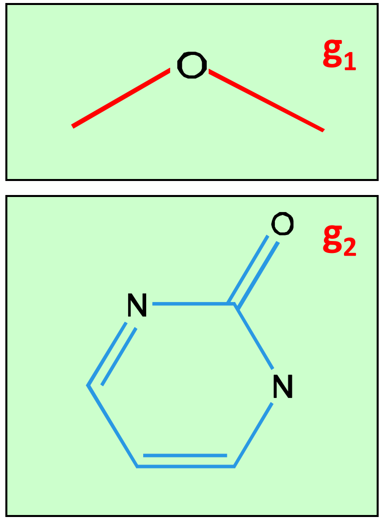
Pattern-Based Classification on Graphs



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 or Pattern-Based Classification

- **Data:** Transactions, microarray data, … → **Patterns or association rules**
- **Classification Methods** (Some interesting work):
  - CBA [Liu, Hsu & Ma, KDD'98]: Use high-conf., high-support *class association rules* to build classifiers — To be discussed here
  - Emerging patterns [Dong & Li, KDD'99]: Patterns whose support changes significantly between the two classes
  - CMAR [Li, Han & Pei, ICDM'01]: Multiple rules in prediction — To be discussed here
  - CPAR [Yin & Han, SDM'03]: Beam search on multiple prediction rules
  - RCBT [Cong et al., SIGMOD'05]: Build classifier based on mining top-k covering rule groups with row enumeration (for high-dimensional data)
  - Lazy classifier [Veloso, Meira & Zaki, ICDM'06]: For a test t, project training data D on t, mine rules from D_t, predict on the best rule
  - Discriminative pattern-based classification [Cheng et al., ICDE'07] — To be discussed here

Advanced Pattern-Based Classification

- Classification: Basic Concepts
- Pattern-Based Classification
- Associative Classification
- Discriminative Pattern-Based Classification
- Direct Mining of Discriminative Patterns
- DPClass: Effective but Concise Discriminative Patterns-Based Classification

Thanks to Hong Cheng@CUHK and Jingbo Shang @UIUC for their contributions

CBA: Classification Based on Associations

- CBA [Liu, Hsu and Ma, KDD'98]
- Method
  - 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; o.w. apply the default rule
  - Effectiveness: Often found more accurate than some traditional classification methods, such as C4.5
  - Why? — Exploring high confident associations among multiple attributes may overcome some constraints introduced by some classifiers that consider only one attribute at a time

CMAR: Classification Based on Multiple Association Rules

- Rule pruning whenever a rule is inserted into the tree
  - Given two rules, R_1 and R_2, if the antecedent of R_1 is more general than that of R_2 and $conf(R_1) \geq conf(R_2)$, 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

Advanced Pattern-Based Classification

- Classification: Basic Concepts
- Pattern-Based Classification
- Associative Classification
- Discriminative Pattern-Based Classification
- Direct Mining of Discriminative Patterns
- DPClass: Effective but Concise Discriminative Patterns-Based Classification

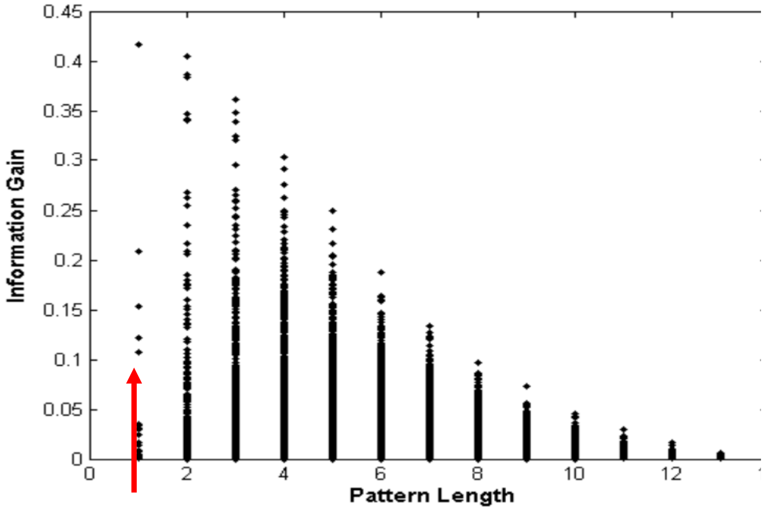
Thanks to Hong Cheng@CUHK and Jingbo Shang @UIUC for their contributions

Discriminative Pattern-Based Classification

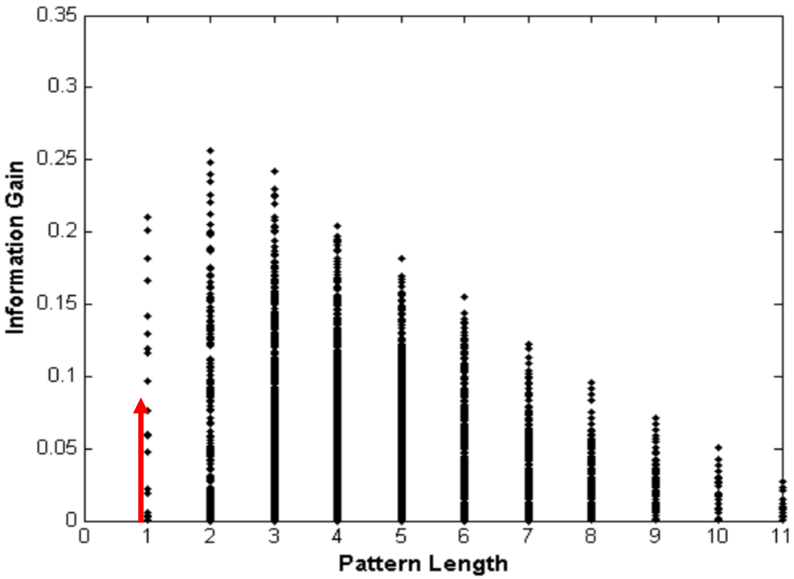
- Discriminative patterns as features for classification [Cheng et al., ICDE'07]
- **Principle:** Mining discriminative frequent patterns as high-quality features and then apply any classifier
- **Framework (PatClass)**
  - Feature construction by *frequent itemset mining*
  - Feature selection (e.g., 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

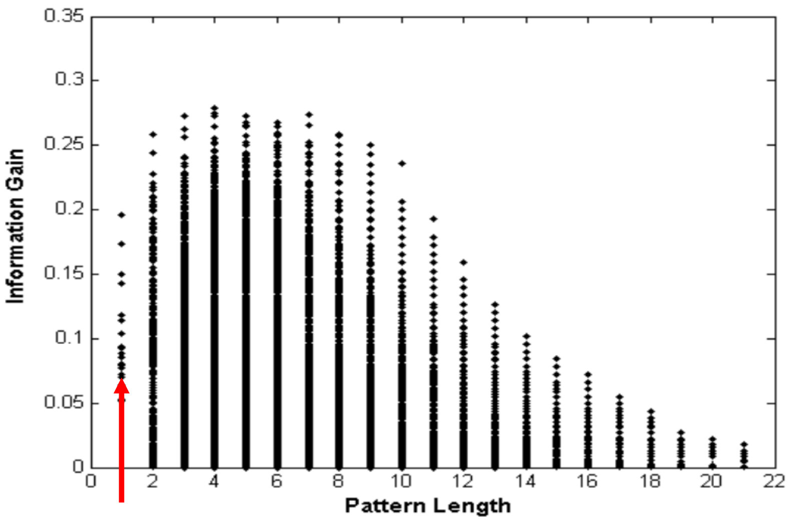
- K-itemsets are often more informative than single features (1-itemsets) in classification
- Computation on real datasets shows: The discriminative power of k-itemsets (for k > 1 but often ≤ 10) is higher than that of single features



(a) Austral



(b) Cleve

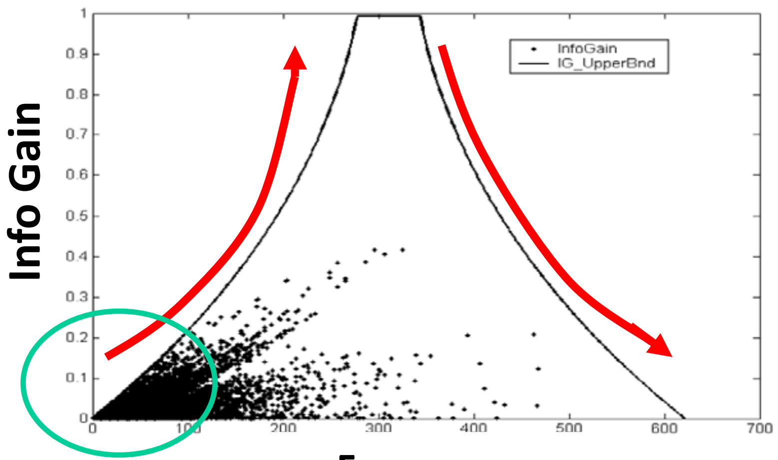


(c) Sonar

Information Gain vs. Pattern Length

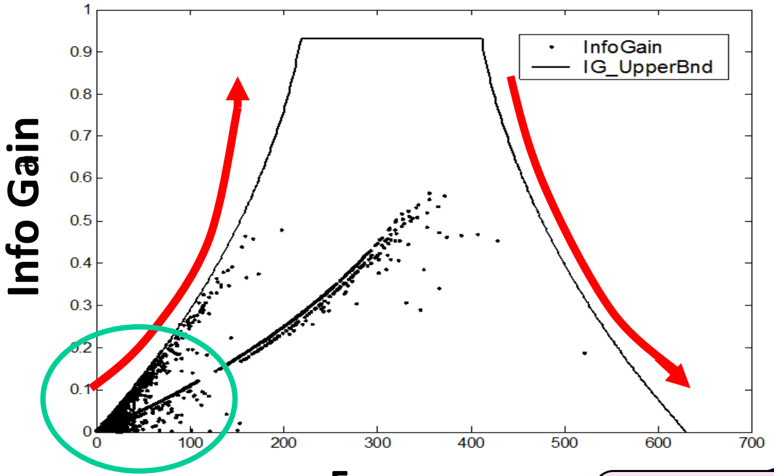
Information Gain vs. Pattern Frequency

- 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

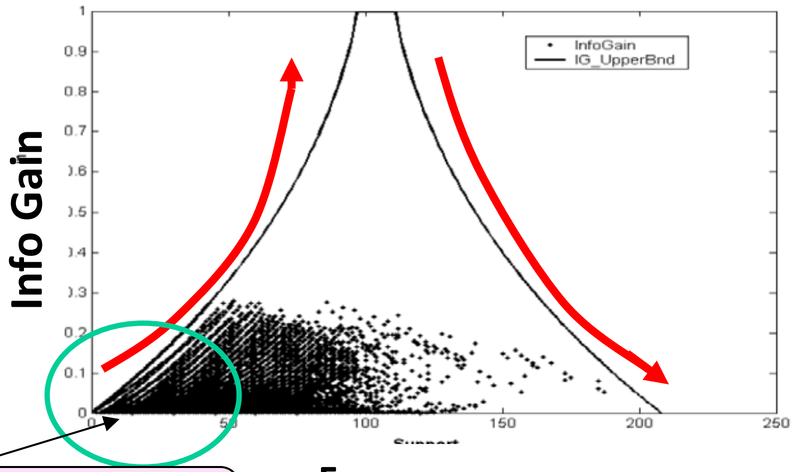
(a) Austral



Frequency

(b) Breast

Low support, low info gain



Frequency

(c) Sonar

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

Entropy of given data

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

Conditional entropy of study focus

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

Discriminative Pattern-Based Classification: Experimental Results

Table 1. Accuracy by SVM on Frequent Combined Features vs. Single Features

Data	Single Feature			Freq. Pattern	
	Item_All	*Item_FS*	*Item_RBF*	*Pat_All*	*Pat_FS*
anneal	**99.78**	**99.78**	99.11	99.33	99.67
austral	85.01	85.50	85.01	81.79	**91.14**
auto	83.25	84.21	78.80	74.97	**90.79**
breast	97.46	97.46	96.98	96.83	**97.78**
cleve	84.81	84.81	85.80	78.55	**95.04**
diabetes	74.41	74.41	74.55	77.73	**78.31**
glass	75.19	75.19	74.78	79.91	**81.32**
heart	84.81	84.81	84.07	82.22	**88.15**
hepatic	84.50	89.04	85.83	81.29	**96.83**
horse	83.70	84.79	82.36	82.35	**92.39**
iono	93.15	94.30	92.61	89.17	**95.44**
iris	94.00	**96.00**	94.00	95.33	**96.00**
labor	89.99	91.67	91.67	94.99	**95.00**
lymph	81.00	81.62	84.29	83.67	**96.67**
pima	74.56	74.56	76.15	76.43	**77.16**
sonar	82.71	86.55	82.71	84.60	**90.86**
vehicle	70.43	72.93	72.14	73.33	**76.34**
wine	98.33	99.44	98.33	98.30	**100**
zoo	97.09	97.09	95.09	94.18	**99.00**

Table 2. Accuracy by C4.5 on Frequent Combined Features vs. Single Features

Dataset	Single Features		Frequent Patterns	
	Item_All	*Item_FS*	*Pat_All*	*Pat_FS*
anneal	98.33	98.33	97.22	**98.44**
austral	84.53	84.53	84.21	**88.24**
auto	71.70	77.63	71.14	**78.77**
breast	95.56	95.56	95.40	**96.35**
cleve	80.87	80.87	80.84	**91.42**
diabetes	**77.02**	**77.02**	76.00	76.58
glass	75.24	75.24	76.62	**79.89**
heart	81.85	81.85	80.00	**86.30**
hepatic	78.79	85.21	80.71	**93.04**
horse	83.71	83.71	84.50	**87.77**
iono	92.30	92.30	92.89	**94.87**
iris	**94.00**	**94.00**	93.33	93.33
labor	86.67	86.67	**95.00**	91.67
lymph	76.95	77.62	74.90	**83.67**
pima	75.86	75.86	76.28	**76.72**
sonar	80.83	81.19	**83.67**	**83.67**
vehicle	70.70	71.49	**74.24**	73.06
wine	95.52	93.82	96.63	**99.44**
zoo	91.18	91.18	95.09	**97.09**

Discriminative Pattern-Based Classification: Scalability Tests

Table 3. Accuracy & Time on Chess Data

min_sup	#Patterns	Time (s)	SVM (%)	C4.5 (%)
1	N/A	N/A	N/A	N/A
2000	68,967	44.703	92.52	97.59
2200	28,358	19.938	91.68	97.84
2500	6,837	2.906	91.68	97.62
2800	1,031	0.469	91.84	97.37
3000	136	0.063	91.90	97.06

Table 4. Accuracy & Time on Waveform Data

min_sup	#Patterns	Time (s)	SVM (%)	C4.5 (%)
1	9,468,109	N/A	N/A	N/A
80	26,576	176.485	92.40	88.35
100	15,316	90.406	92.19	87.29
150	5,408	23.610	91.53	88.80
200	2,481	8.234	91.22	87.32

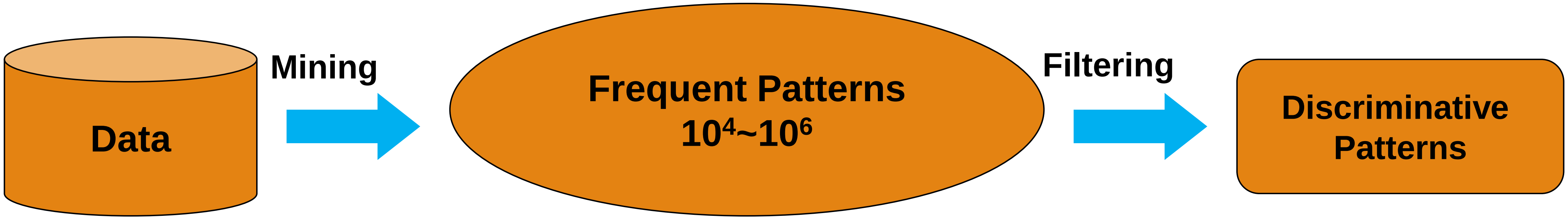
Advanced Pattern-Based Classification

- Classification: Basic Concepts
- Pattern-Based Classification
- Associative Classification
- Discriminative Pattern-Based Classification
- Direct Mining of Discriminative Patterns
- DPClass: Effective but Concise Discriminative Patterns-Based Classification

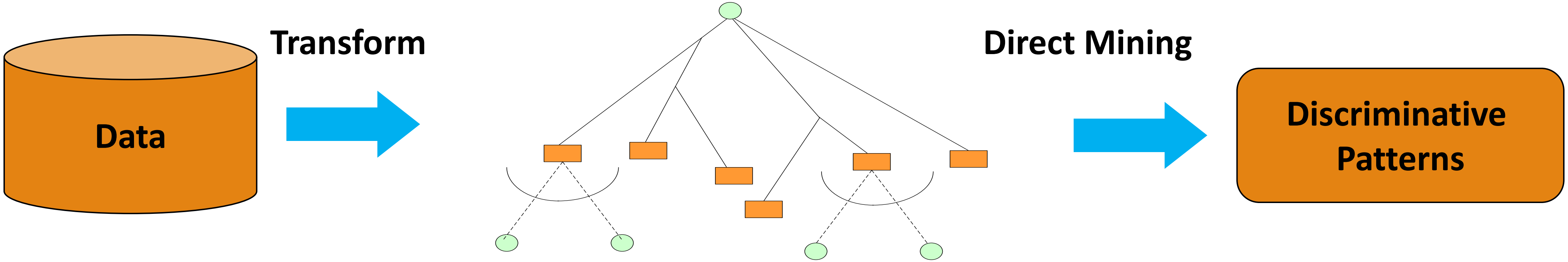
Thanks to Hong Cheng@CUHK and Jingbo Shang @UIUC for their contributions

Direct Mining of Discriminative Patterns

Frequent pattern mining, then getting discriminative patterns: Expensive



Direct mining of discriminative patterns : Efficient



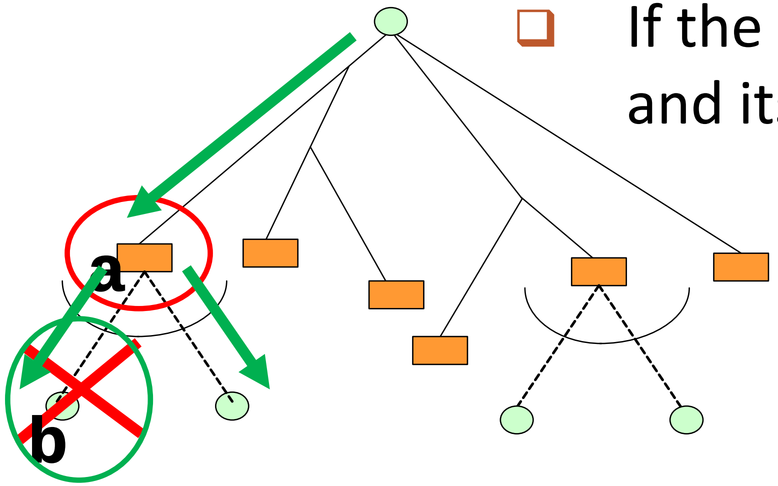
FP-tree

DDPMine: Direct Discriminative Pattern Mining

- DDPMine [Cheng et al., ICDE'08]: Efficient, direct discriminative pattern mining
- **General methodology**
  - Input: A set of training instances D and a set of features F
  - Iteratively perform feature selectin based on the "sequential coverage" paradigm
    - Select the feature f_i with the highest discriminative power
    - Remove instances D_i from D covered by the selected feature f_i
- **Implementation**
  - Integration of branch-and-bound search with FP-growth mining
  - Iteratively eliminate training instances and progressively shrink the FP-tree

DDPMine: Branch-and-Bound Search

- The discriminative power (information gain) of a low frequency pattern is upper bounded by a small value
- During FPGrowth mining we record the most discriminative itemset discovered so far and its information gain value g_{best}
  - Before constructing a conditional FP-tree, we first estimate the upper bound of information gain based on the conditional DB
  - If the upper bound value ≤ g_{best}, skip this conditional FP-tree and its subsequent trees

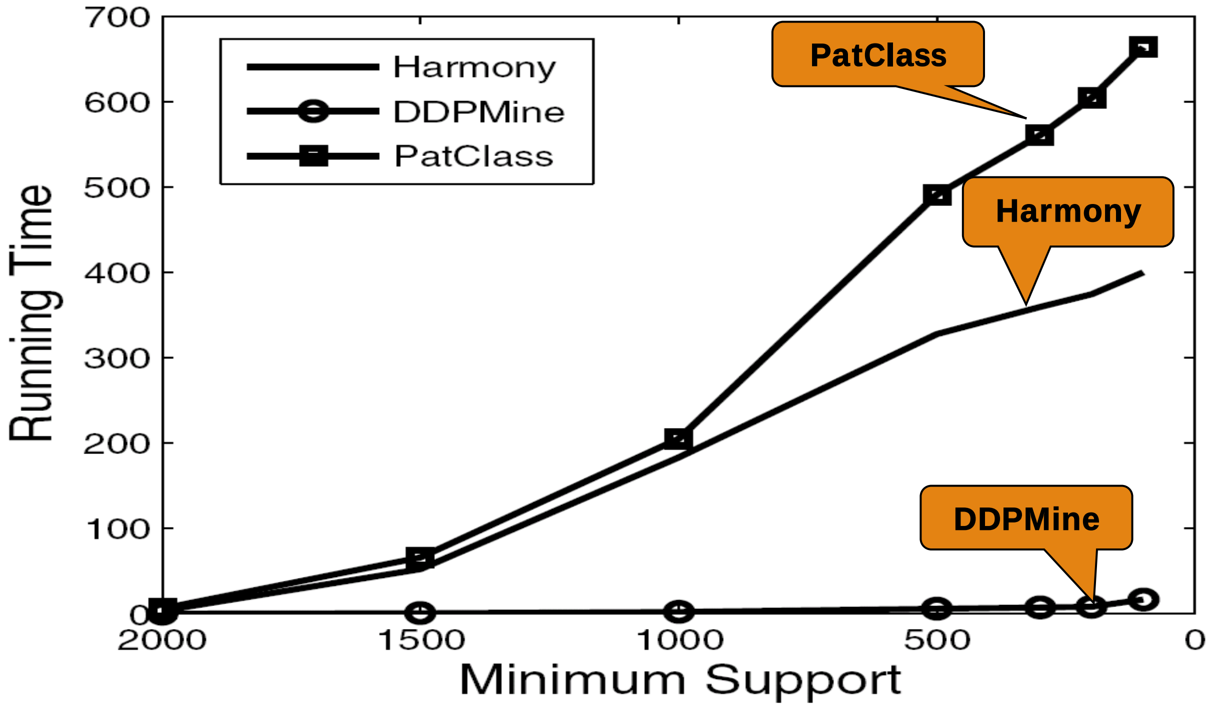


Upper bound-based FP-tree pruning

- Ex.: Prune b's cond. FP-tree if UpperBoundIG(b) ≤ InfoGain(a), where UpperBound IG(b) is determined by b's support in its conditional DB
- DDPMine: A feature-based approach, i.e., mining only the most discriminative patterns

DDPMine Efficiency: Runtime Comparison

- Comparing three algorithms on classification efficiency (runtime in seconds)
  - PatClass: Discriminative-Pattern-Based Classification [Cheng et al., ICDE'07]
  - Harmony [Wang & Karypis, SDM'05]
  - DDPMine: Direct discriminative pattern mining [Cheng et al., ICDE'08]



- All three methods mine discriminative frequent patterns for effective classification
- DDPMine substantially improves mining efficiency

A Comparison on Classification Accuracy

- In comparison with Harmony and PatClass, DDPMine maintains high accuracy and substantially improves mining efficiency
- An extension of this methodology has been applied to software bug analysis (D. Lo, et al., "Classification of Software Behaviors for Failure Detection: A Discriminative Pattern Mining Approach", KDD'09

Datasets	Harmony	PatClass	DDPMine
adult	81.90	84.24	84.82
chess	43.00	91.68	91.85
crx	82.46	85.06	84.93
hypo	95.24	99.24	99.24
mushroom	99.94	99.97	100.00
sick	93.88	97.49	98.36
sonar	77.44	90.86	88.74
waveform	87.28	91.22	91.83
Average	82.643	92.470	92.471

Advanced Pattern-Based Classification

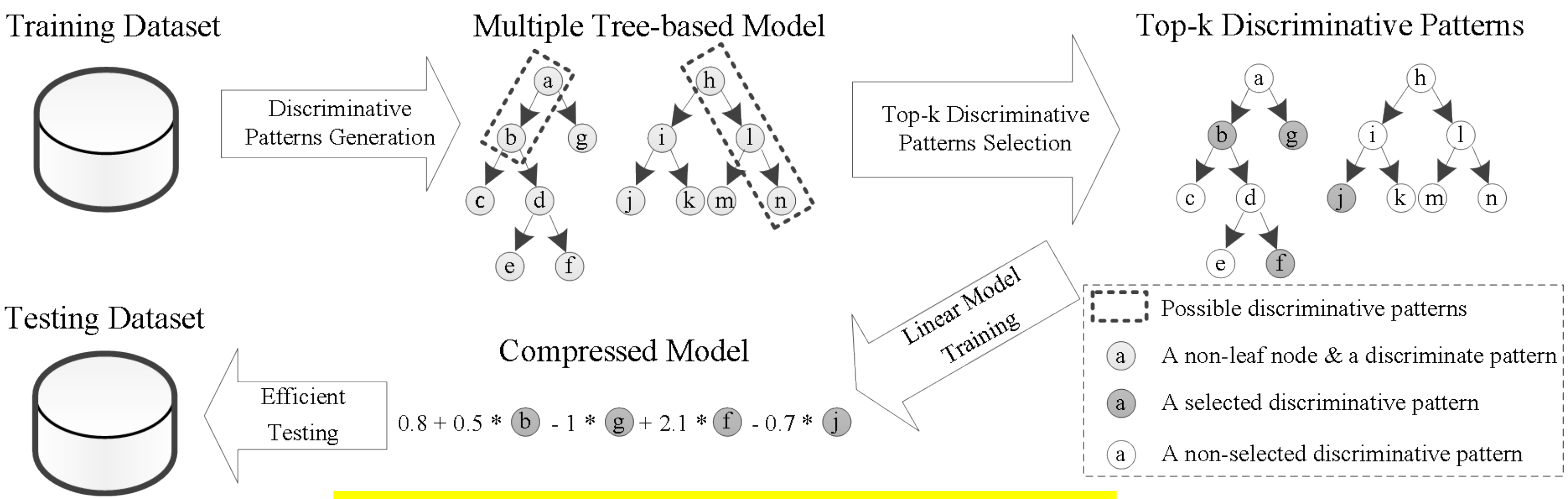
- Classification: Basic Concepts
- Pattern-Based Classification
- Associative Classification
- Discriminative Pattern-Based Classification
- Direct Mining of Discriminative Patterns
- DPClass: Effective but Concise Discriminative Patterns-Based Classification

Thanks to Hong Cheng@CUHK and Jingbo Shang @UIUC for their contributions

Why DPClass?—Concerns over Previous Models

- Single tree models, e.g., decision tree/boosted tree
  - **Sensitive** to training instances → **overfitting**
- Multiple trees models, e.g., random forest
  - **Tree-independent**: the growth & traditional pruning strategies
  - Model size could be **very large** → **slow** online prediction
  - **Uninterpretable**
- Problems of frequent and discriminative pattern methods: PatClass and DDPMine
  - Frequent does not necessarily imply discriminative
  - The number of frequent patterns might be very large
  - This may imply a large but useless pool of frequent patterns

DPClass: Compatible Discriminative Patterns for Linear Models



An Overview of the DPClass Framework

- Train a constrained multiple tree-based model
  - Discriminative pattern: Every prefix path from the root of a tree to any of its non-leaf nodes
- Two solutions to select top-*k* discriminative patterns

Discriminative and Top-K Patterns

- Discriminative patterns: Strong signals on the specific classification task
  - E.g., a pattern with very high information gain
- Top-k Patterns
  - Top-*k* patterns: A *size-k subset of discriminative patterns, which have the best performance (i.e., the accuracy in classification tasks) based on the training data*
  - Some effects of different patterns may have a large portion of overlaps, e.g. $(v_0 \cap v_1 \cap v_2)$ and $(v_0 \cap v_1 \cap v_2 \cap v_3)$
  - A set of patterns is compatible $\triangleq$ They have strong signals on the specific classification task and every single pattern has its own "significant" contributions

Generation of Discriminative Patterns

- DPClass: A binary classification task
  - N training instances $(x_1, y_1), (x_2, y_2) \dots (x_N, y_N), \forall 1 \le i \le N, y_i \in \{+1, -1\}$
    - x_i is the feature vector of i-th instance
  - Both numeric (continuous) and categorical (discrete) variables are acceptable
- Step 1: Generation of discriminative patterns: Based on Random Forest
  - Maximize the randomness
    - Random features, random partitions, random instances (bootstrap)
  - Parameters: # of trees = T; loss function = information gain; depth $\le d$; support $\ge \sigma$ (based on bootstrapped instances)
  - We admit all prefix of these tree-paths as patterns
    - # of leaves$\le \min\left\{2^d, \frac{N}{\sigma}\right\} \cdot T$; # of candidate patterns $\le \min\left\{2^d, \frac{N}{\sigma}\right\} \cdot T \cdot d$
  - Assume $T = 100$, # of candidate pattern $\sim 10^4$

Selection of Compatible Discriminative Patterns

- Select a k-set of most compatible discriminative patterns
- Implementation
  - Forward Selection (Greedy)
  - LASSO (GLMNET)

Experiments: On Machine Learning Repository Data

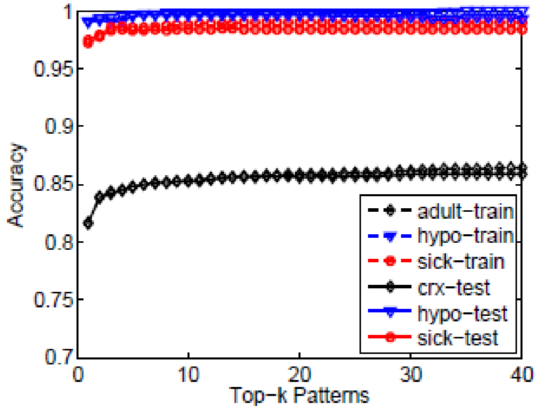
Classification accuracy (when k = 20) vs. RF (Random Forest without any constraints) and DDPMine

Dataset	adult	hypo	sick	crx	sonar	chess	namao	musk	madelon
DPClass-F	**85.66%**	**99.58%**	98.35%	**89.35%**	**85.29%**	92.25%	97.17%	95.92%	74.50%
DPClass-L	84.33%	99.28%	**98.87%**	87.96%	83.82%	92.05%	96.94%	95.71%	**76.00%**
RF	85.45%	97.22%	94.03%	**89.35%**	83.82%	**94.22%**	**97.86%**	**96.60%**	56.50%
DDPMine	83.42%	92.69%	93.82%	87.96%	73.53%	90.04%	96.83%	93.29%	59.83%

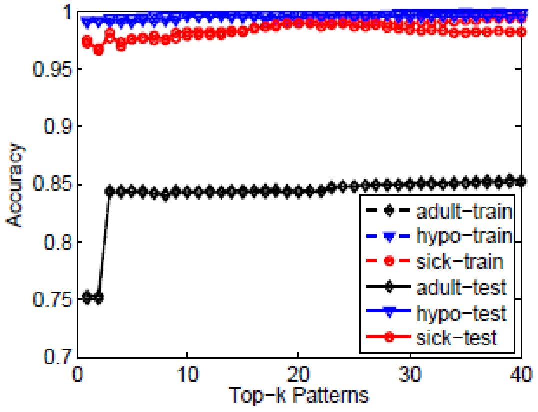
DDPMine outperforms decision tree and support vector machine on all these UCI Machine Learning datasets

The impact of top-k patterns. Training and testing accuracies are almost overlapped.

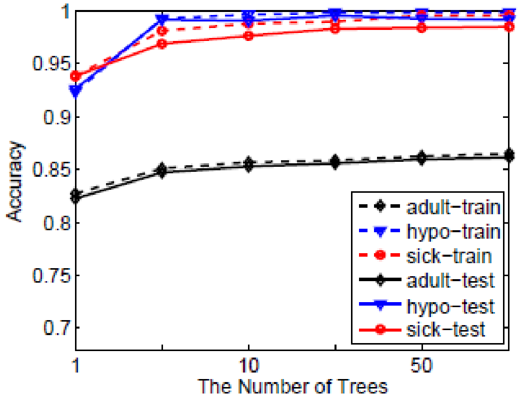
The impact of the number of trees. Training and testing accuracies are almost overlapped.



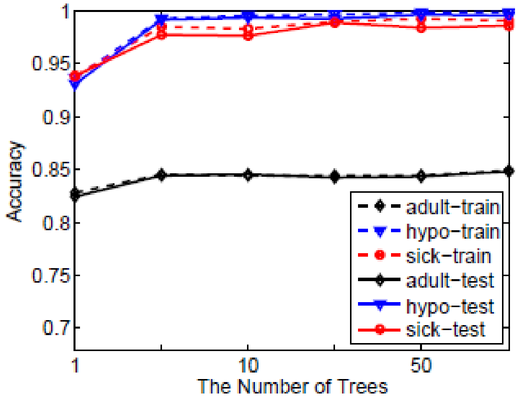
(a) DPClass-F



(b) DPClass-L



(a) DPClass-F



(b) DPClass-L

Summary on DPClass

- DPClass can compress the model and thus the online prediction is extremely fast
- DPClass have comparable performance as advanced models
  - Even better in experiments
- DPClass can learn the interpretable patterns
- Extensible to other discriminate patterns learning tasks
  - Ex.: Multi-class classification, regression, and survival analysis

Recommended Readings

- H. Cheng, X. Yan, J. Han, C.-W. Hsu, Discriminative Frequent Pattern Analysis for Effective Classification, ICDE'07
- H. Cheng, X. Yan, J. Han, P. S. Yu, Direct Discriminative Pattern Mining for Effective Classification, ICDE’08
- G. Cong, K. Tan, A. Tung & X. Xu. Mining Top-k Covering Rule Groups for Gene Expression Data, SIGMOD’05
- M. Deshpande, M. Kuramochi, N. Wale & G. Karypis. Frequent Substructure-based Approaches for Classifying Chemical Compounds, TKDE’05
- G. Dong & J. Li. Efficient Mining of Emerging Patterns: Discovering Trends and Differences, KDD’99
- W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. S. Yu & O. Verscheure. Direct Mining of Discriminative and Essential Graphical and Itemset Features via Model-based Search Tree, KDD’08
- W. Li, J. Han & J. Pei. CMAR: Accurate and Efficient Classification based on Multiple Class-association Rules, ICDM’01
- B. Liu, W. Hsu & Y. Ma. Integrating Classification and Association Rule Mining, KDD’98
- J. Shang, W. Tong, J. Peng, and J. Han. DPClass: An Effective but Concise Discriminative Patterns-Based Classification Framework, SDM'16
- J. Wang and G. Karypis. HARMONY: Efficiently Mining the Best Rules for Classification, SDM’05
- X. Yin & J. Han. CPAR: Classification Based on Predictive Association Rules, SDM’03