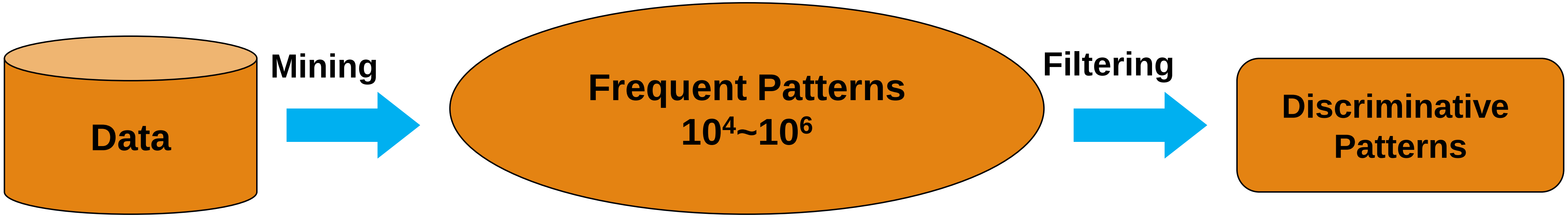
Session 5. DDPMine: Direct Mining of Discriminative Patterns

Direct Mining of Discriminative Patterns

Frequent pattern mining, then getting discriminative patterns: Expensive



Data

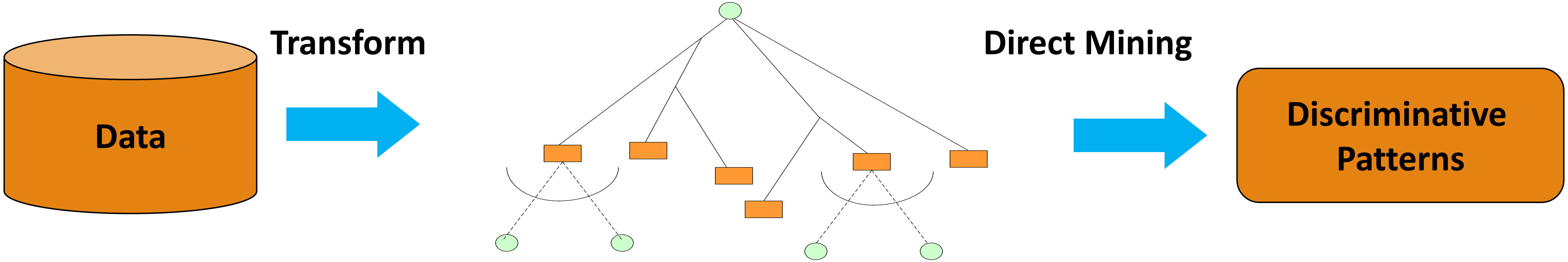
Mining

Frequent Patterns
10^4~10^6

Filtering

Discriminative Patterns

Direct mining of discriminative patterns : Efficient



Data

Transform

Direct Mining

Discriminative Patterns

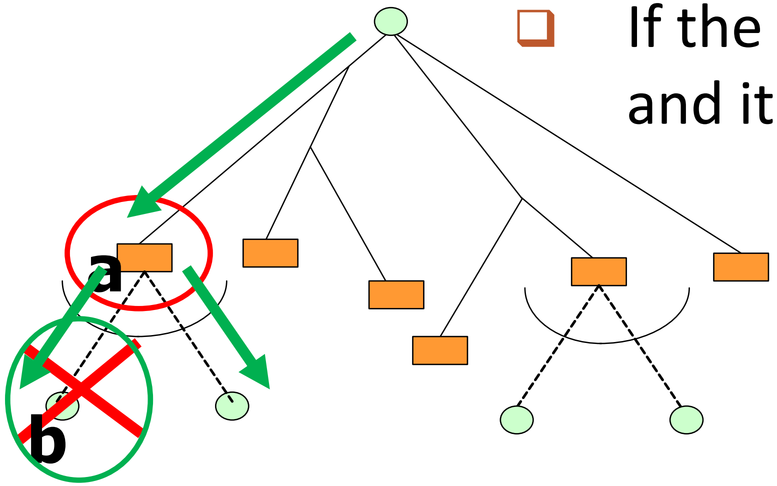
FP-tree

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

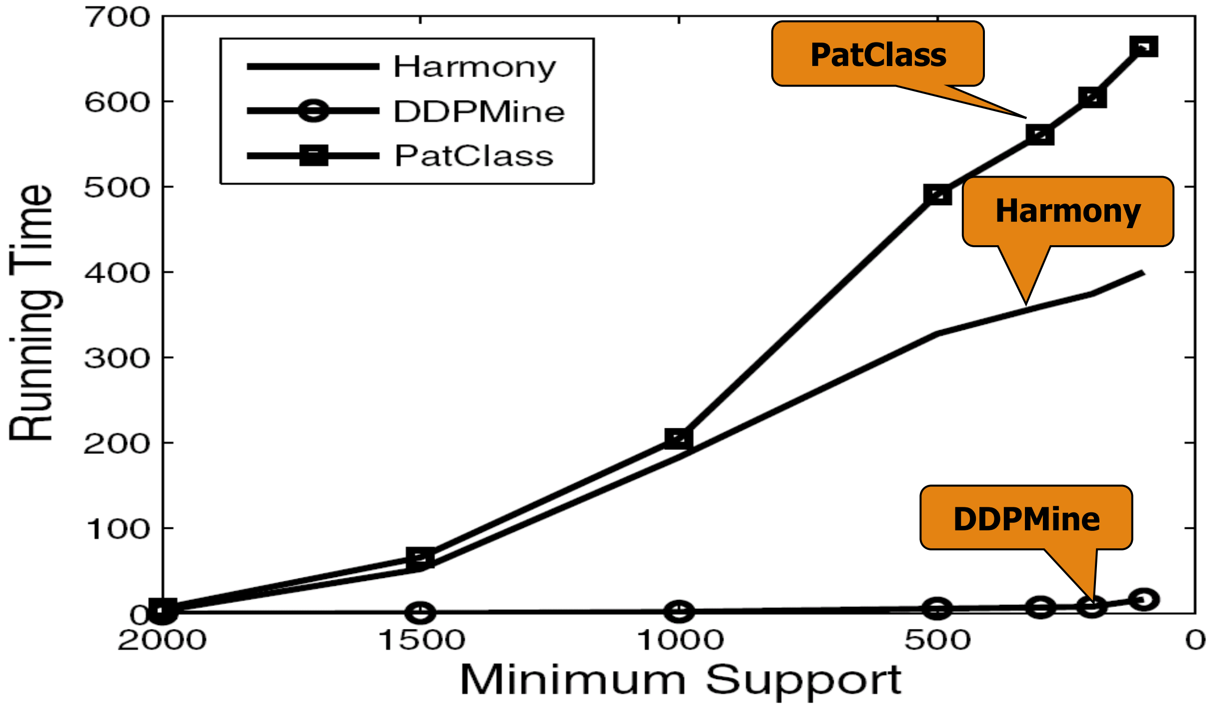
- The discriminative power (information gain) of a low frequency pattern is upper bounded by a small value
- During FPGrowth mining we record the most discriminative itemset discovered so far and its information gain value g_{best}
    - Before constructing a conditional FP-tree, we first estimate the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



Upper bound-based FP-tree pruning

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

Summary

- Concepts of classification and pattern-based classification
- Associative classification methods, such as CBA and CMAR
- Discriminative pattern-based classification
- Direct mining of discriminative patterns: DDPMine

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. 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