Frequent Pattern-based Classification and Post-Processing of Mining Results

Hong Cheng

Data Mining Group

University of Illinois at Urbana-Champaign

DAIS The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign

Large Scale Information Management

Part I: Frequent Pattern-based Classification

Basic Idea

- Mine discriminative frequent patterns;
- Represent the data in the feature space of such patterns;
- Build classification models.

Application

- Transactional database
  - Relational dataset, Customer transaction data, etc.
  - Frequent itemsets
- Sequence database
  - Protein sequences, Web log data, etc.
  - Frequent sequential patterns or K-substrings
- Graph database
  - Chemical compounds, Molecules, etc.
  - Frequent substructures

Frequent pattern is a good candidate as features, especially for data with complicated structures.

Why Are Frequent Patterns Useful?

- Frequent pattern
  - A non-linear combination of single features
  - Increase the expressive power of the feature space
    - Exclusive OR example
    - Data is linearly separable in (x, y, xy), but not in (x, y)

X	Y	C
0	0	0
0	1	1
1	0	1
1	1	0

Mining

Transform

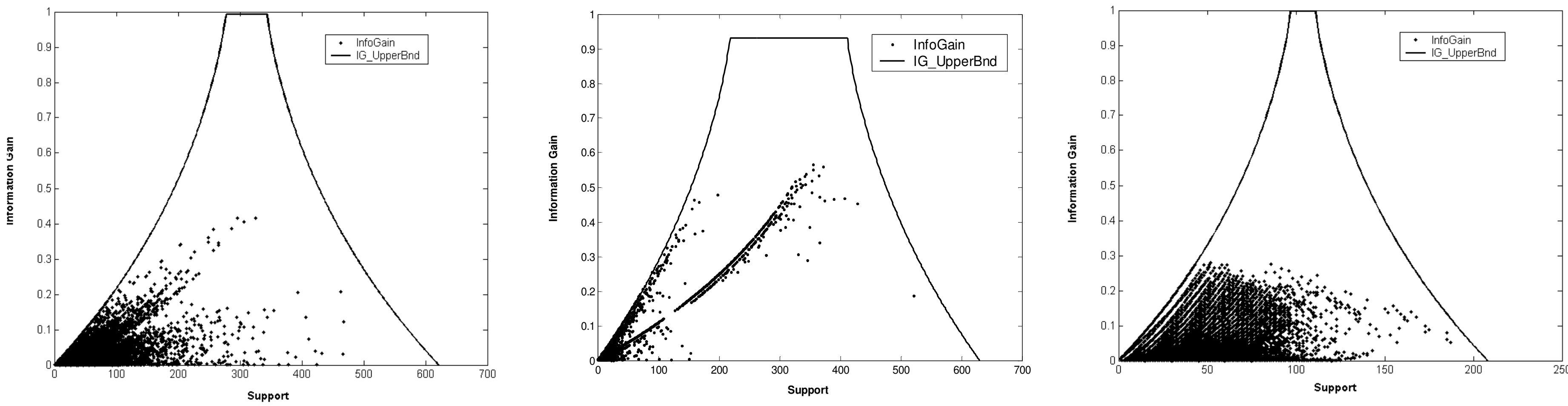
X	Y	XY	C
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

Linear classifier $f((x, y, xy)^T) = x + y - 2xy$

Discriminative Power vs. Frequency

- The discriminative power of a feature is closely related to its frequency.
- The discriminative power of a low-frequency feature is low!
- Theoretical analysis [Cheng et al, ICDE'07]

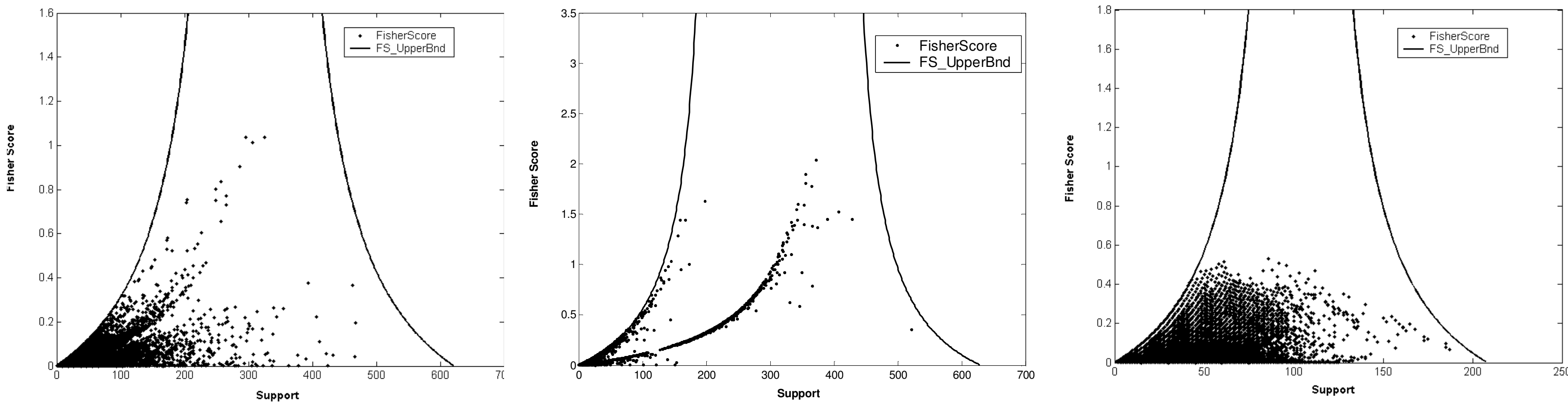
Information Gain vs. Frequency



(a) Austral (b) Breast (c) Sonar

Fig. 1. Information Gain vs. Pattern Frequency

Fisher Score vs. Frequency



(a) Austral (b) Breast (c) Sonar

Fig. 2. Fisher Score vs. Pattern Frequency

Experimental Results

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

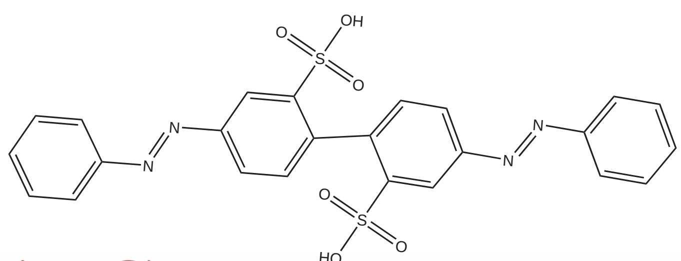
Data	Single Feature			Freq. Pattern	
	Item_All	*Item_FS*	*Item_RBF*	*Pat_All*	*Pat_FS*
anneal	**99.78**	**99.78**	99.11	99.33	99.67
austral	85.01	85.50	85.01	81.79	**91.14**
auto	83.25	84.21	78.80	74.97	**90.79**
breast	97.46	97.46	96.98	96.83	**97.78**
cleve	84.81	84.81	85.80	78.55	**95.04**
diabetes	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**

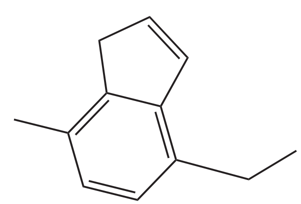
Table 2. Accuracy by C4.5 on Frequent Combined Features vs. Single Features

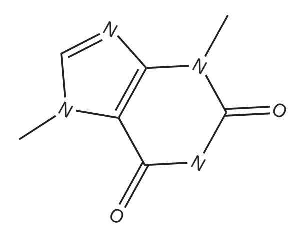
Dataset	Single Features		Frequent Patterns	
	Item_All	*Item_FS*	*Pat_All*	*Pat_FS*
anneal	98.33	98.33	97.22	**98.44**
austral	84.53	84.53	84.21	**88.24**
auto	71.70	77.63	71.14	**78.77**
breast	95.56	95.56	95.40	**96.35**
cleve	80.87	80.87	80.84	**91.42**
diabetes	**77.02**	**77.02**	76.00	76.58
glass	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**

Graph Classification

- A learning approach to assign class labels (toxic/non-toxic, active/inactive) to graph data such as molecules or chemical compounds.
- Applications
  - *QSAR* in chemical informatics
  - Screening in drug design







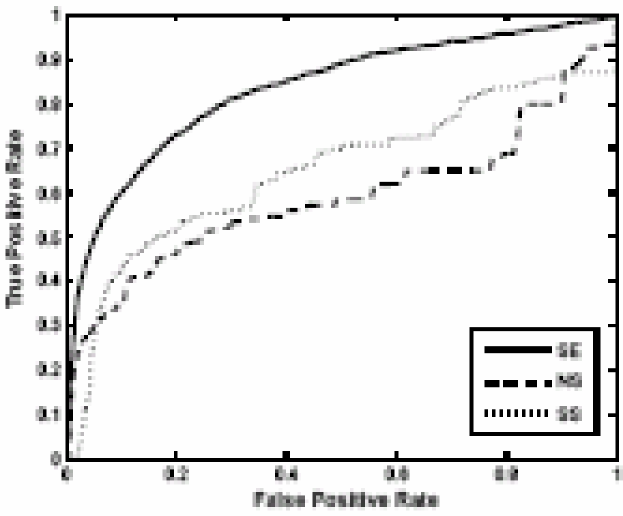
Challenges in Graph Classification

- Feature construction and selection
  - Data not in readily available feature vector format
  - Simple features such as atoms or edges not discriminative
  - Structural features are better candidates
- Skewed class distribution
  - AIDS anti-viral screen datasets
    - Active class : only 1%
  - NCI anti-cancer screen datasets
    - Active class : around 5%

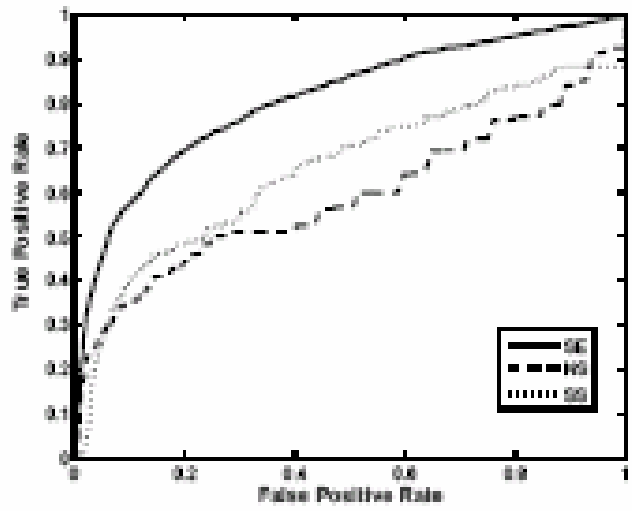
An Ensemble Approach

- Structural features
  - Discriminative frequent subgraphs
- Sampling
  - Repeated samples of the positive class
  - Under samples of the negative class
- Ensemble
  - Build multiple classifiers based on different balanced data samples
  - Reduce the variance introduced by sampling

ROC Plot



(a) ROC, NCI1



(b) ROC, NCI81

Experimental Results

Table 4: ROC50, Base Learner C4.5

Datasets	SE	SE+FE	GF
NCI1	0.4880	**0.5279**	0.3260
NCI109	0.4361	**0.5909**	0.3020
NCI123	**0.4853**	0.4808	0.2630
NCI145	0.5235	**0.5887**	0.3400
NCI167	0.5047	**0.5715**	0.0640
NCI33	0.4419	**0.5175**	0.3180
NCI330	0.5183	**0.5687**	0.3430
NCI41	0.4392	**0.5362**	0.3570
NCI47	**0.4987**	0.4971	0.3110
NCI81	0.4252	**0.4689**	0.2950
NCI83	0.5152	**0.5761**	0.3170
H1	0.4655	**0.5956**	0.2680
H2	0.3960	0.6059	**0.6510**

Part II: Post-Processing of Mining Results

From Mining to Understanding and Application

Frequent Itemsets → Association Rules → Sequential Patterns → Graph Patterns →

bottleneck: huge number of patterns

Applications:
indexing, classification, prediction, clustering

Post-processing of Mining Results

- Pattern Summarization [Yan et al, KDD’05]
  - Pattern compression with a maximal preservation of pattern and support information by exploring pattern profiles
  - Won Best Student Paper Runner-up Award
- Pattern Compression [Xin et al, VLDB’05]
  - Find a set of representative patterns which can cover the rest of patterns with bounded distance
- Top-K Pattern Extraction [Xin et al, KDD’06]
  - Pick the most important K patterns
  - Avoid picking redundant patterns
- Semantic Annotation [Mei et al, KDD’06]
  - Annotate a frequent pattern with in-depth, concise and structured information
  - Won Best Student Paper Runner-up Award

Thank You

hcheng3@uiuc.edu

www.ews.uiuc.edu/~hcheng3