Pattern Spaces: Theory, Techniques, & Applications

PI: Wong Limsoon, National University of Singapore

- Objectives:
  - Theoretical properties of pattern spaces
  - Algo for their mining
  - Algo for their incremental maintenance
  - Ways to build classifiers based on them
- Novelty:
  - Patterns with more complex measures
  - Dynamic aspects of pattern spaces
  - Improve link with statistics
- Scope & Deliverables:
  - Understanding of structural properties of pattern spaces
  - Algo for mining such pattern spaces & their compact reps
  - Algo for maintaining such compact reps when underlying db changes
  - Accurate classifiers based on such patterns

- Team Members:


Wong Limsoon (PI)


Feng Mengling (RA)


Lee Terk Shuen (Student)


Liu Guimei (RF)


Ngo Thanh Son (RA)


Wang Yue (Student)

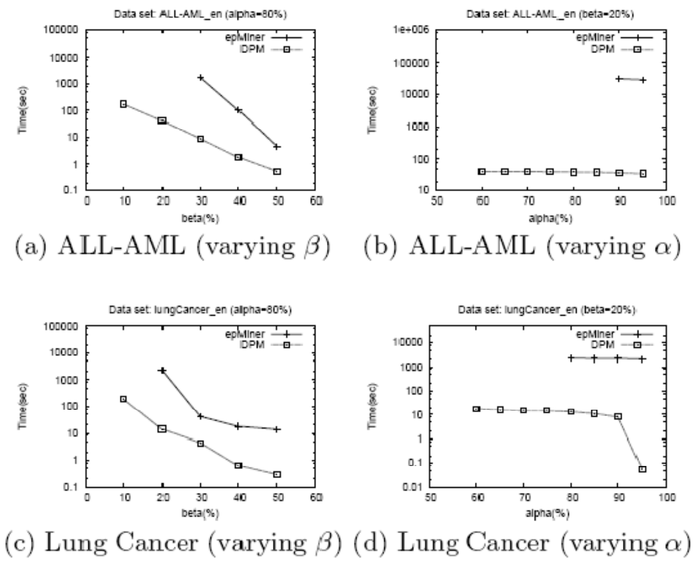

Donny Soh (RA)


Wilson Goh (RA)

- Achievement #1
  - Decomposition of various pattern spaces as convex equiv classes
  - Equiv classes & positive borders as compact rep of various pattern spaces
  - Efficient simultaneous mining of equiv classes of patterns having good odds ratio, relative risk, $\chi 2$ & other statistics

- Associated Technologies
  - Fast algo (DPMiner, GrGrowthPBd) for mining of equiv classes & positive borders satisfying a variety of statistics

- References
  - J. Li, et al. **Mining Statistically Important Equivalence Classes and Delta-Discriminative Emerging Patterns.** *Proc. KDD 2007,* pages 430—439
  - G. Liu, et al. **A New Concise Representation of Frequent Itemsets Using Generators and a Positive Border**. *KAIS*, 17:35—56, 2008



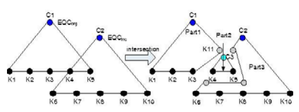
(a) ALL-AML (varying β) (b) ALL-AML (varying α)

(c) Lung Cancer (varying β) (d) Lung Cancer (varying α)

Figure 5: Running time comparison between DP-Miner and epMiner.

α = min freq threshold in +ve samples, β = max freq theshold in –ve samples

- Achievement #2
  - Full understanding of structural changes to pattern equiv classes as underlying db evolves ➔ Exact characterization of equiv classes that emerge, disappear, split, or merge



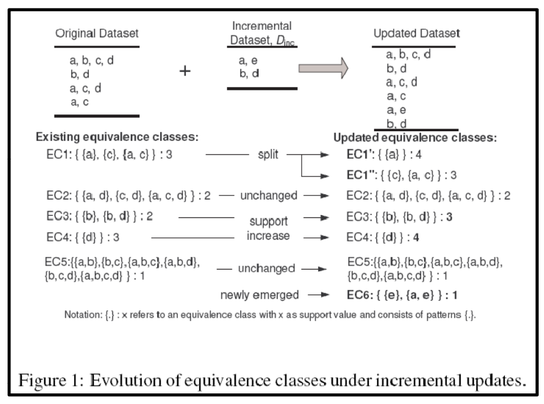
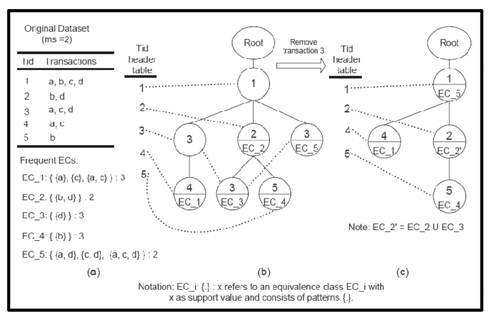


Figure 1: Evolution of equivalence classes under incremental updates.

- Reference
  - M. Feng, et al**. Pattern Space Maintenance for Data Updates & Interactive Mining**. *Comput Intel*, 26(3):282--317, August 2010

- Achievement #3
  - Efficient maintenance of pattern equiv classes when transactions are removed
- Associated Technology
  - Fast algo (TRUM) for trend analysis of insert-only db
  - Novel data structure Tid-tree



Notation: EC_i {.} : x refers to an equivalence class EC_i with x as support value and consists of patterns {.}.

- Reference
  - M. Feng et al. **Evolution and Maintenance of Frequent Pattern Space when Transactions are Removed**. *Proc. PAKDD 2007*, pages 489--497

- Achievement #4
  - Efficient maintenance of pattern equiv classes when transactions are removed/ added or threshold is changed
- Associated Technology
  - Fast algo (PSM) for interactive pattern mining & trend analysis

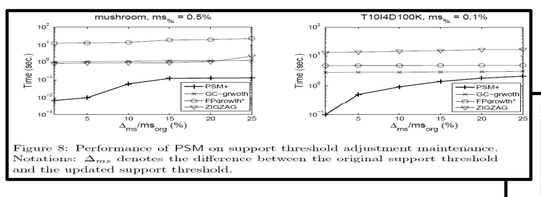


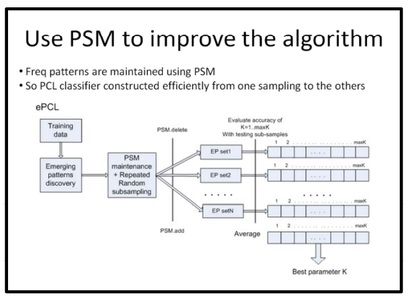
Figure 8: Performance of PSM on support threshold adjustment maintenance. Notations: Δ_{ms} denotes the difference between the original support threshold and the updated support threshold.

dataset	#PSM+	#FPgrowth*	#GC-growth
bms-pos ($ms_\% = 0.1\%$)	80	110K	110K
bms-webview1 ($ms_\% = 0.1\%$)	250	3K	3K
chess ($ms_\% = 40\%$)	350K	6M	1M
connect-4 ($ms_\% = 20\%$)	80K	1800M	1M
mushroom ($ms_\% = 0.5\%$)	10K	300M	165K
pumb* ($ms_\% = 30\%$)	2K	400K	27K
retail ($ms_\% = 0.1\%$)	270	8K	8K
T10I4D100K ($ms_\% = 0.5\%$)	11	1K	1K
T10I10D100K ($ms_\% = 10\%$)	7K	70K	55K

Table 1: Comparison of the number of patterns enumerated by PSM+, FP-growth* and GC-growth. Notations: #PSM+, #FPgrowth* and #GC-growth denote the approximated number of patterns enumerated by the respectively algorithms.

- References
  - M. Feng et al. **Negative Generator Border for Effective Pattern Maintenance**. *Proc. ADMA 2008*, pages 217--228
  - M. Feng, et al**. Pattern Space Maintenance for Data Updates & Interactive Mining**. *Comput Intel*, 26(3):282--317, August 2010

- Achievement #5
  - Statistically sound choice of # of patterns to use in robust pattern-based classifiers
  - Fast sampling (using incremental pattern maintenance) for efficiently applying Central Limit Theorem
- Associated Technology
  - Robust classifier (ePCL) based on emerging pattern generators



- Reference
  - Ngo et al. **Efficiently Finding the Best Parameter for the Emerging Pattern-based Classifier PCL**. *Proc. PAKDD 2010*

Contact- Professor Wong Limsoon (wongls@comp.nus.edu.sg)

Project URL- http://www.comp.nus.edu.sg/~wongls/projects/pattern-spaces/index.html

Funding- A*STAR SERC PSF grant # 072 101 0016

NUS National University of Singapore
School of Computing