Discriminative Pattern Mining

Alexandre Termier, Lacodam

HyAIAI meeting @ home

07/05/2020

Prelude: a quick pattern mining refresher

Alice				
Bob				
Charles				
Diana				
Erin				

Input:

- *Transactional dataset D*
- *Minimum support value (ex: minsup = 2)*

Output:

all subsets P of {, , , } s.t. P appears in at least 2 transactions of D

- Frequent itemsets:
  - {, , } support : 2
  - {, } support: 2
  - {, } support: 2
  - {, } support: 3
  - {, } support: 3
  - {, } support: 2
  - {} support: 4
  - {} support: 3
  - {} support: 4
  - {} support: 3

Introduction

- Grand goal of pattern mining: find useful/meaningful patterns
  - Totally unsupervised case: this is hard!

- Some data come with hints on interest: multi-class datasets
  - **Dual-class**: Disease / Not disease, Poisonous / Edible, Spam / Not spam
  - Multi-class: Young / Adult / Old, US / UK / FR / JP…

- Discriminative pattern mining:
  - Input: dual-class dataset
  - Find patterns characteristic of a class
    - Also called: *contrast* PM, *emerging* PM

Interest of discriminative pattern mining

- Get better understanding of class
  - Ex: better understand disease (symptoms, affected people, genotype…)
  - Ex: Mushroom data :
    - {odor = none, stalk-surface-below-ring = smooth, ring-number = one} : edible 57%, poisonous 0.2%

- Build (interpretable) classifiers

- Monitoring
  - Increase / decrease of dissimilarity + symptoms
  - Ex: live stream of system measurement versus reference in controlled environment

Applications: spotlight on bioinformatics

- High-order SNP combinations
  - SNP : Single-Nucleotide Polymorphism
  - Correlate groups of SPNs with diseases (or phenotypic traits)
  - Pb: huge number of SNPs (human = 5 millions)

- Differential gene expressions
  - Gene = item, Cell type = transaction
  - Cell can be cancerous or not
  - Value = level of expression of gene for given cell (discretized)
  - Goal : discover groups of genes that are constrained to specific intervals of gene expression

- Regulatory motif combinations
  - Transcriscription factors (TF) -> help cells to respond to various signals
  - Usually response come from groups of TF
  - => find most significant groups of TF for a response

Discriminance measures

Discriminance measures

- Measures to evaluate how much a pattern is characteristic of a class

- Many measures have been proposed in the literature

- Can rely on lots of related work in statistics !

Contingency table

D: complete dataset, 2 classes: 1 and 2

D_1: elements of D of class 1

D_2: elements of D of class 2

	Presence	Absence	Row total
D_1	t_{11}	t_{12}	$\lvert D_1 \rvert = t_{11} + t_{12}$
D_2	t_{21}	t_{22}	$\lvert D_2 \rvert = t_{21} + t_{22}$
Column total	t_1	t_2	$\lvert D \rvert = \lvert D_1 \rvert + \lvert D_2 \rvert$

	1	0	Σ
D_1	t_{11}	t_{12}	$\|D_1\|$
D_2	t_{21}	t_{22}	$\|D_2\|$
Σ	t_1	t_2	$\|D\|$

Basic measures

Given p a pattern:

- Difference of support

$$DS(p, D_1, D_2) = |\ sup(p, D_1) - sup(p, D_2)\ | = |\ t_{11}/|D_1| - t_{12}/|D_2|\ |$$

- Growth rate

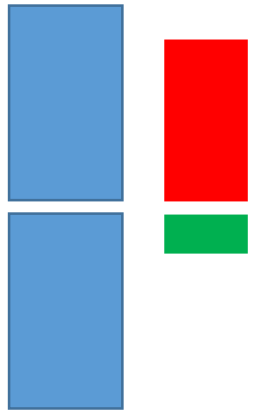
$$GR(p, D_1, D_2) = \frac{\sup(p, D_1)}{\sup(p, D_2)} = \frac{t_{11}/|D_1|}{t_{12}/|D_2|}$$

Testing the basic measures

	1	0	Σ
D_1	t_{11}	t_{12}	$\|D_1\|$
D_2	t_{21}	t_{22}	$\|D_2\|$
Σ	t_1	t_2	$\|D\|$

	1	0	Σ
D_1	8	2	10
D_2	2	8	10
Σ	10	10	20

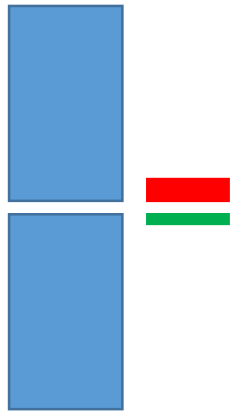
- DS = | 8/10 – 2/10 | = 0.6
- GR = 8 / 2 = 4



Could be significative

	1	0	Σ
D_1	8	392	400
D_2	2	398	400
Σ	10	790	800

- DS = | 8/400 – 2/400 | = 0.015
- GR = 8 / 2 = 4



Real phenomena, or noise ?

Stat. based measures

	1	0	Σ
D_1	t_{11}	t_{12}	$\|D_1\|$
D_2	t_{21}	t_{22}	$\|D_2\|$
Σ	t_1	t_2	$\|D\|$

- Odds ratio

$$OR(p, D_1, D_2) = \frac{t_{11}t_{22}}{t_{12}t_{21}}$$

- Chi square

$$\chi^2 = \sum_{i=1}^{i=2} \sum_{j=1}^{j=2} \frac{(t_{ij} - E_{ij})^2}{E_{ij}}, E_{ij} = \frac{\sum_{q=1}^{q=2} t_{iq} \sum_{q=1}^{q=2} t_{qj}}{|D|}$$

- Mutual Information

$$MI(p, D_1, D_2) = \sum_{i=1}^{i=2} \sum_{j=1}^{j=2} \frac{t_{ij}}{|D|} log \frac{t_{ij}/|D|}{t_i|D_j|/|D|^2}$$

- Information Gain

$$IG(p, D_1, D_2) = sup(p, D_1)(log \frac{sup(p,D_1)}{sup(p,D)} - log \frac{|D_1|}{|D|})$$

Testing measures, part 2

	1	0	Σ
D_1	t_{11}	t_{12}	$\|D_1\|$
D_2	t_{21}	t_{22}	$\|D_2\|$
Σ	t_1	t_2	$\|D\|$

	1	0	Σ
D_1	8	2	10
D_2	2	8	10
Σ	10	10	20

- OR = (8*8) / (2*2) = 16
- X^2 = 7.2
- MI = 0.19
- IG = 9.305

	1	0	Σ
D_1	8	392	400
D_2	2	398	400
Σ	10	790	800

- OR = (8*398 / 2*392) = 4.06
- X^2 = 3.6
- MI = 0.01
- IG = 9.305

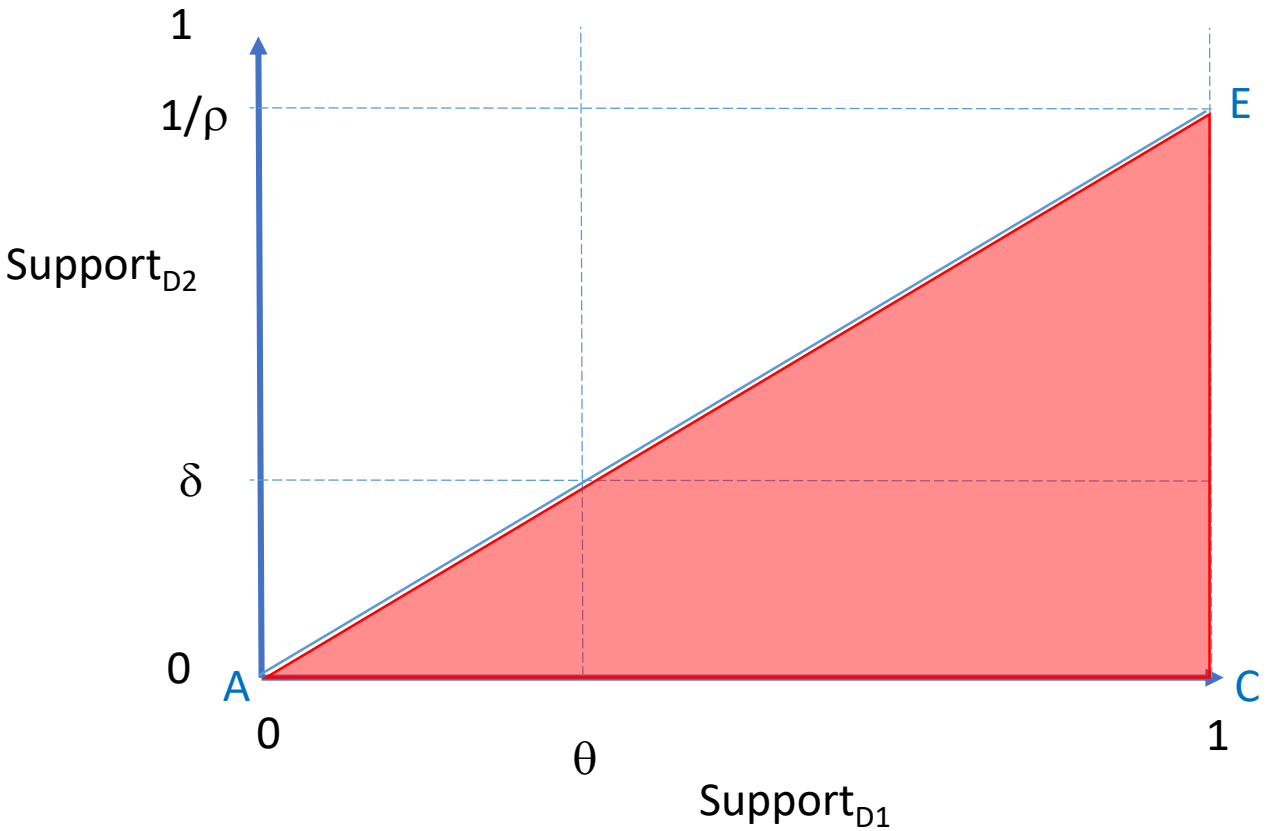
Algorithms

Main problems

- Discriminance measures are not anti-monotonic
  - The discriminance of a pattern does not depend on the discriminance of its parents
  - $\rightarrow$ classical pruning schemes cannot be applied…
- Need a new threshold for the discriminance measure
  - Choosing it correctly is hard

Mining EP with borders [Dong et al, KDD 99]

Setting: discrimance measure = **growth rate**



min sup D1 = θ

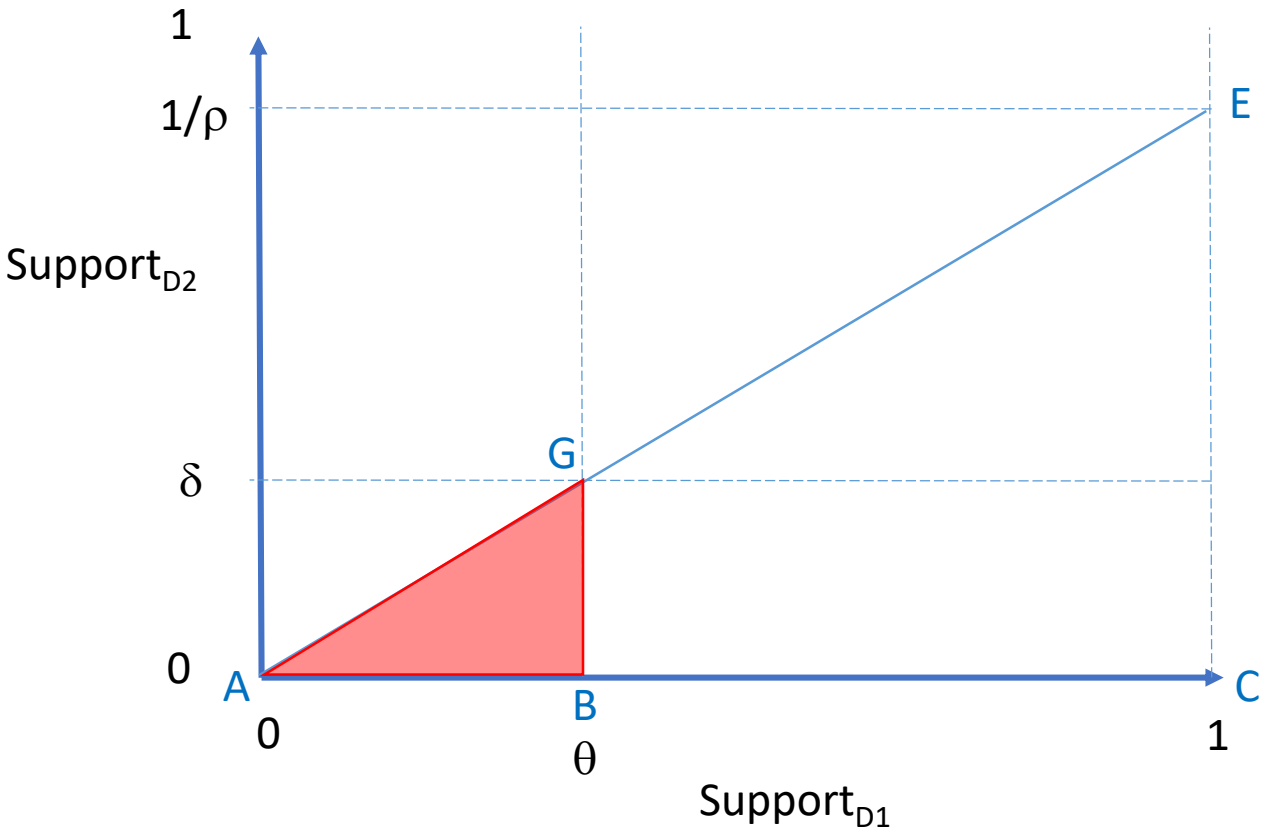
min sup D2 = δ = θ/ρ

GR = $\frac{support_{D1}}{support_{D2}} \geq \rho$

All EP live in the ACE triangle

Mining EP with borders [Dong et al, KDD 99]

Setting: discrimance measure = **growth rate**



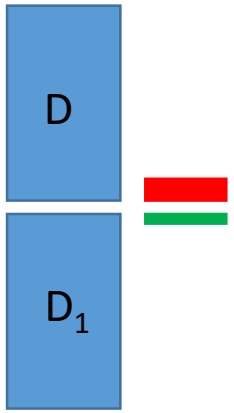
min sup D2 = θ

min sup D1 = δ = θ/ρ

GR = $\frac{support_{D1}}{support_{D2}} \geq \rho$

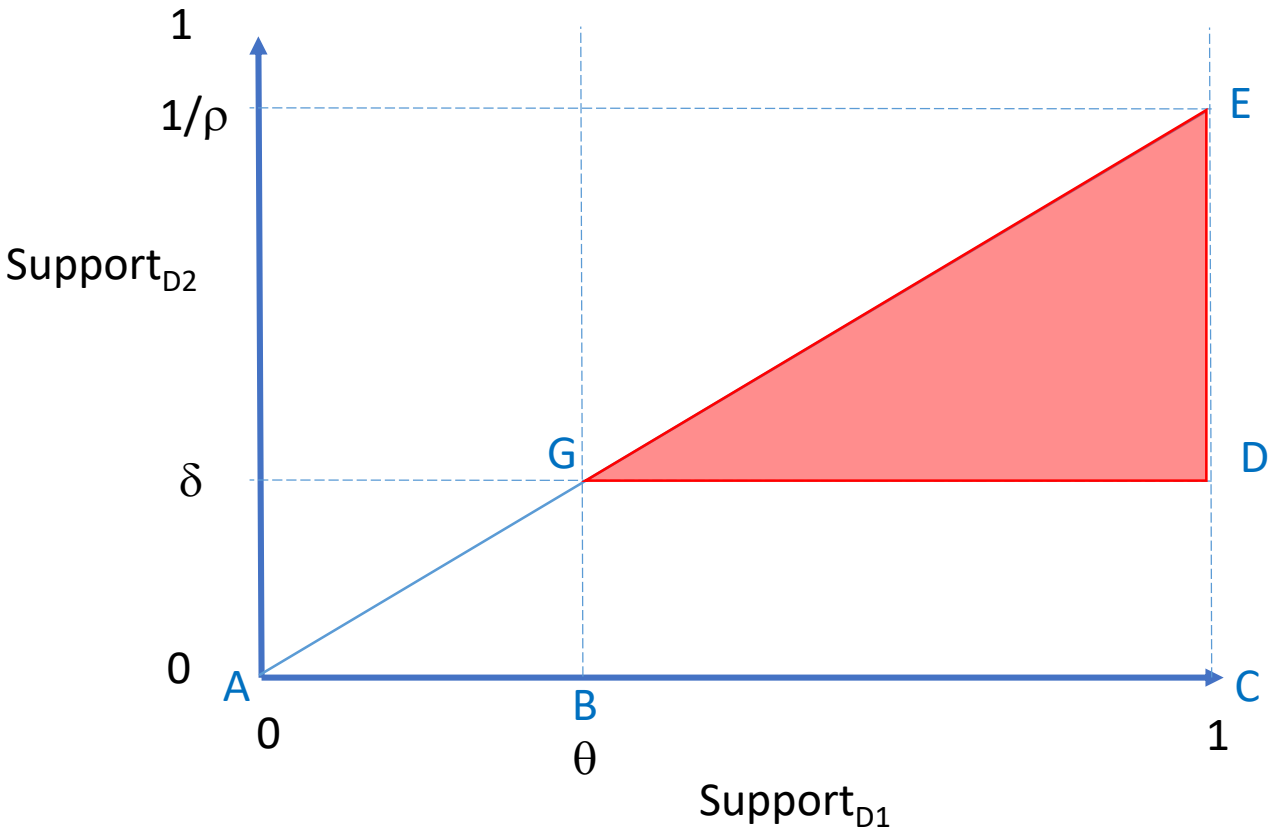
ABG triangle:

- Many Eps
- But low support in both datasets -> hard to compute
- Significance?



Mining EP with borders [Dong et al, KDD 99]

Setting: discrimance measure = **growth rate**



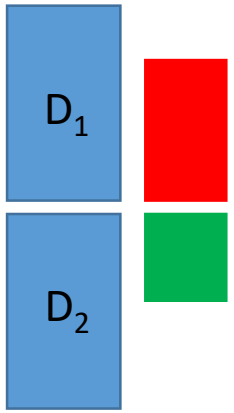
min sup D2 = θ

min sup D1 = δ = θ/ρ

GR = $\frac{support_{D1}}{support_{D2}} \geq \rho$

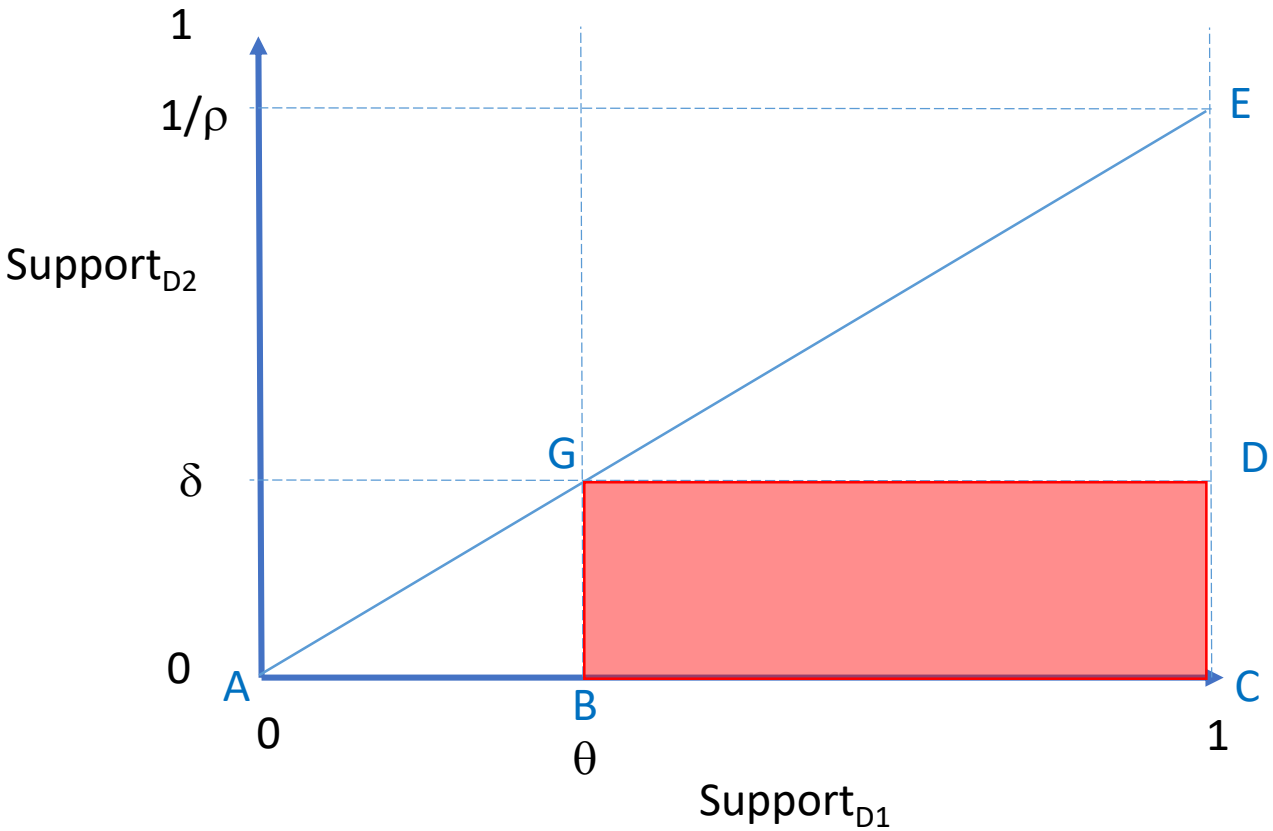
EDG triangle:

- High support both datasets
- -> fewer EPs
- Not the priority to solve
  - Algo in paper



Mining EP with borders [Dong et al, KDD 99]

Setting: discrimance measure = **growth rate**



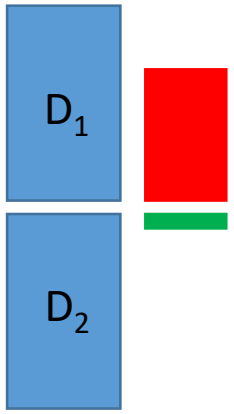
min sup D2 = δ

min sup D1= θ

GR = $\frac{support_{D1}}{support_{D2}} \geq \rho$

BCDG rectangle:

- High support D_2 / low support D_1
- Many *promising* EPs
- Not easy to solve -> KDD 99 algo



Other discriminative measures

- Previous algorithm: designed for Growth Rate

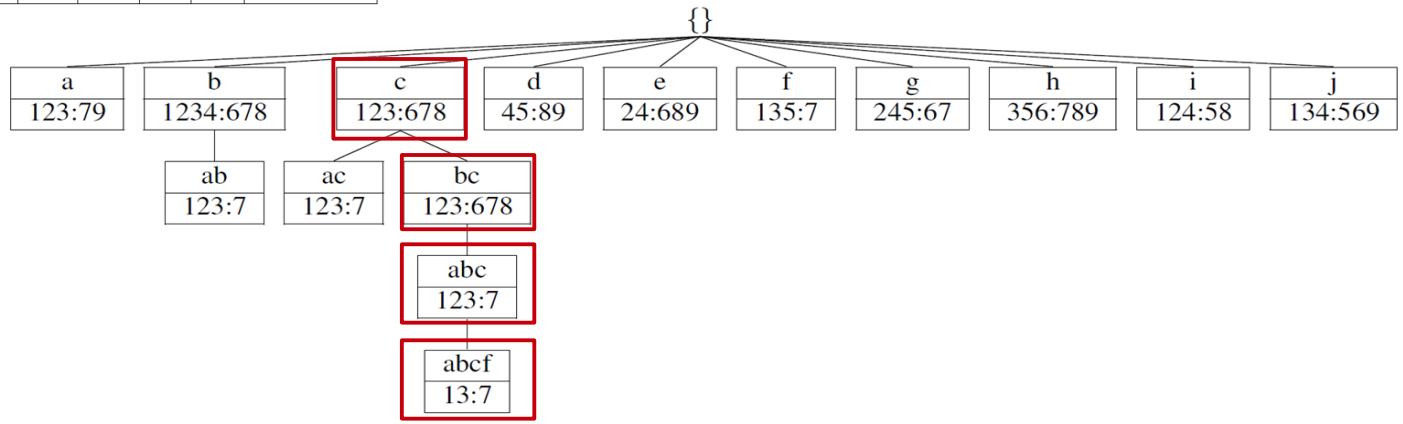
- Other measures?

- I present SSDPS, an algorithm we made for OR and RR
- Designed for bioinformatics data:
  - Many items
  - Few transactions

Hoang-Son Pham, Gwendal Virlet, Dominique Lavenier, Alexandre Termier: Statistically Significant Discriminative Patterns Searching. DaWaK 2019: 105-115

Classical enumeration strategy

Transaction ids	Items										Class
1	a	b	c			f			i	j	1
2	a	b	c		e		g		i		1
3	a	b	c			f		h		j	1
4		b		d	e		g		i	j	1
5				d		f	g	h	i	j	1
6		b	c		e		g	h		j	0
7	a	b	c			f	g	h			0
8		b	c	d	e			h	i		0
9	a			d	e		g	h		j	0



Pruning strategies ?

Itemset	Frequency	Risk score
c	*6*	*OR = 0.5*
bc	*6*	*OR = 0.5*
abc	*4*	*OR = 4.5*
abcf	*3*	*OR = 2.0*
Anti-monotonic ?	***YES***	***NO***

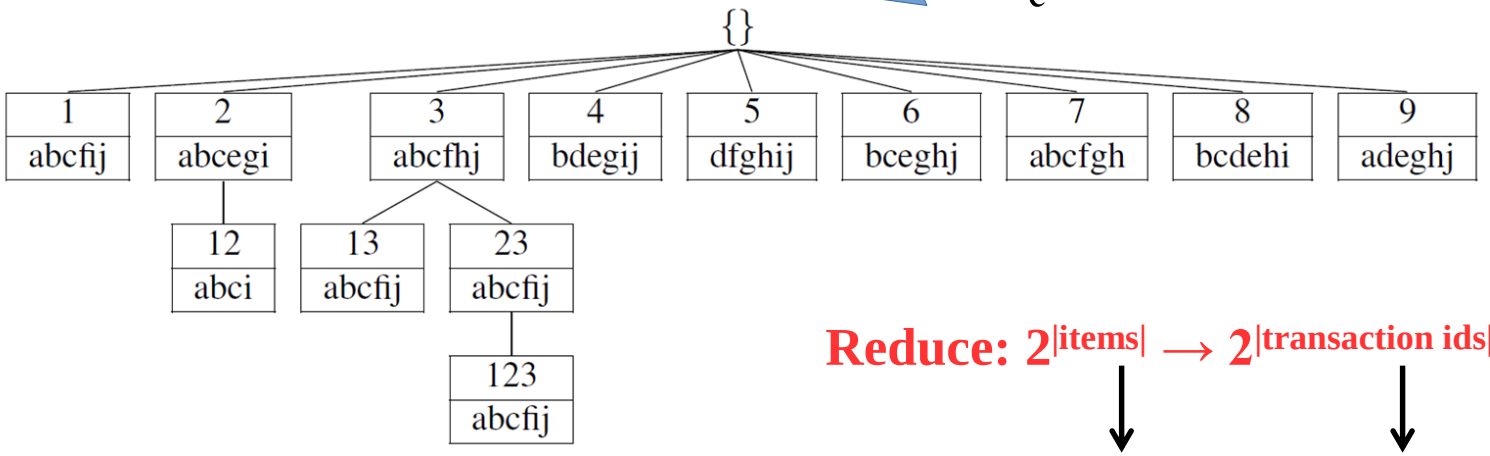
Enumeration on transposed matrix

Transaction ids	Items										Class
1	a	b	c			f			i	j	1
2	a	b	c		e		g		i		1
3	a	b	c			f		h		j	1
4		b		d	e		g		i	j	1
5				d		f	g	h	i	j	1
6		b	c		e		g	h		j	0
7	a	b	c			f	g	h			0
8		b	c	d	e			h	i		0
9	a			d	e		g	h		j	0

Transposition

Items	Transaction ids								
a	1	2	3				7		9
b	1	2	3	4		6	7	8	
c	1	2	3			6	7	8	
d				4	5			8	9
e		2		4		6		8	9
f	1		3		5		7		
g		2		4	5	6	7		9
h			3		5	6	7	8	9
i	1	2		4	5			8	
j	1		3	4	5	6			9
class	1	1	1	1	1	0	0	0	0

Enumerate

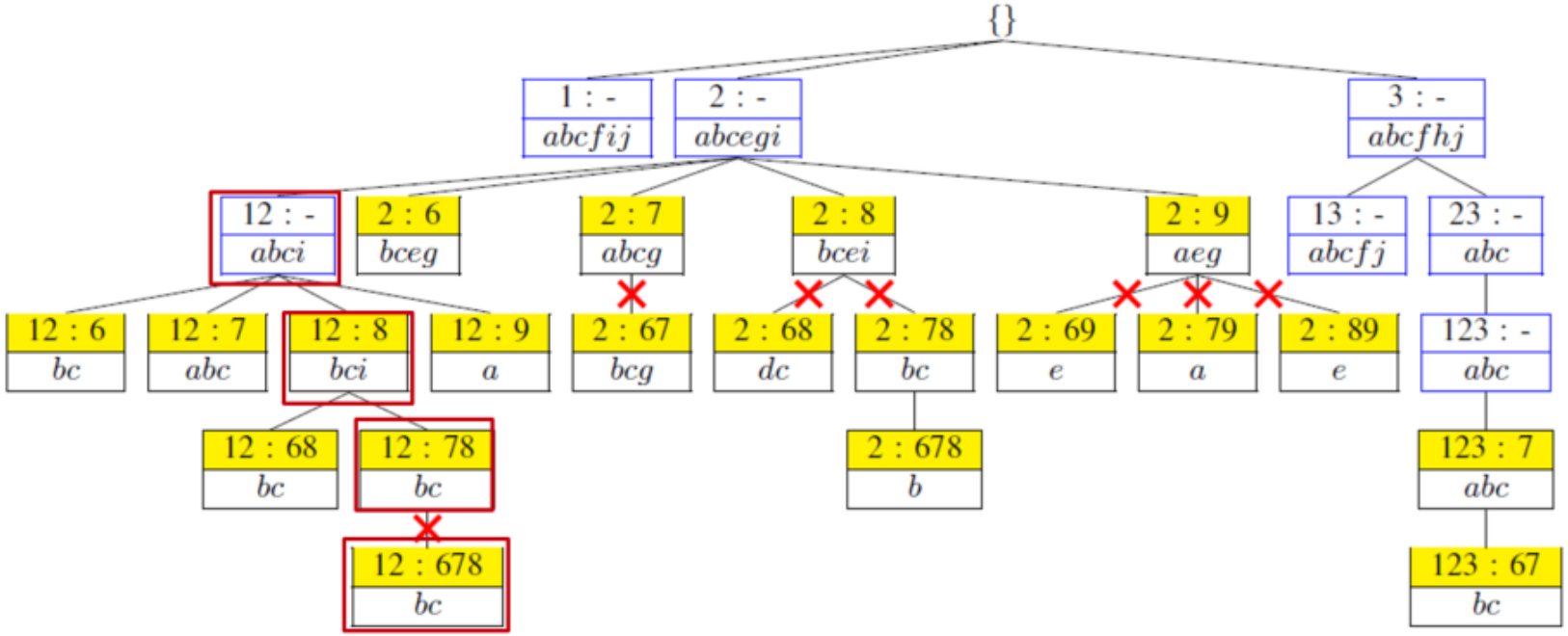


Reduce: $2^{|\text{items}|} \rightarrow 2^{|\text{transaction ids}|}$

10^6 10^3

Some anti-monotonicity returns!

For OR, RR: anti-monotonic on a branch of enumeration tree of transposed matrix



Threshold = 1	*Tidset*	*Itemset*	*Risk score*
-	12 : -	abci	OR = +∞
YES	12 : 8	bci	OR = 2*3 / 3*1 = 2
NO	12 : 78	bc	OR = 2*2 / 3*2 = 0.66
Pruned	12 : 678	bc	OR = 2*1 / 3*3 = 0.22

Statistical significance

Some statistics

- Previous measures give us some info on how discriminative patterns can be
- But does it have statistical meaning?

- $\rightarrow$ need to compute **statistical significance**
  - p-value
  - confidence interval

Definitions

- **p-value**
  - Test to determine if null hypothesis can be rejected or not
    - Here null hypothesis is: *the pattern is not discriminant*
  - p-value = Proba(current pattern occurences | null-hypothesis is true)
  - If p-value < 0.05, then null hypothesis can be rejected
    - This only means that the pattern is unlikely to come from noise
    - At most 5% False Positives with this value

- Confidence Interval (CI)
  - Determine a confidence interval [LCI, UCI] for a statistic measure (ex: Odds Ratio - OR)
    - OR = 1 means that the pattern is not characteristic of a class
    - If 1 in [LCI, UCI] then null hypothesis cannot be rejected
    - Here also threshold (usually 95%)

Multiple hypothesis testing

- If $N = 2^{|Items|}-1$ patterns, then N p-value tests should be made
  - *Hence « multiple hypothesis testing »*

- But (at most) 5% false positives with significance level at 0.05
  - N is huge so large number of false positives, and we don't know which ones!
  - FWER (Family Wise Error Rate) = proba of at least one False Positive

- Solution: make corrections to the significance level to guarantee false positive rate

Control of FWER

- Bonferroni correction
  - Parameters: K nb of tests to do, α significance level (0.05)
  - Method: For all tests, reject null hypothesis only if p-value < α / K
  - Pb:
    - K = nb of patterns to test – unknown !
    - If setting K = $2^{|items|}$ -1, becomes ridiculously strict

- LAMP (Terada et al, PNAS 2013)
  - Very infrequent patterns should not be counted as hypothesis to test
  - Non-closed patterns should not be counted as hypothesis to test
  - Allows a better counting of hypothesis -> better calibration of Bonferroni correction

Conclusion

- Discriminative pattern mining = good tool to discover patterns relevant to a class
- Can be used to build (interpretable) classifiers
- Problem of error correction: how far can it be ignored?
- Still output too many patterns in many cases
  - « Dirty » solution (biologists): put (many) statistical filters for post-processing
  - « Clean » solution (data miners):
    - Patterns sets of discriminative patterns…
    - …with MDL (DiffComp algorithm, group of J. Vreeken)