Pattern mining under preferences

Engelbert Mephu Nguifo
LIMOS – UBP -- UMR 6158 CNRS

December 10, 2014 -- LATECE Seminar, UQAM, Montreal

Université Blaise Pascal (Clermont II)

http://www.univ-bpclermont.fr

- Pluridisciplinaire, 16007 étudiants en 2013/2014,

- 5 Unités de Formation et de Recherche (UFR) :
  - UFR Langues Appliquées, Commerce et Communication : LACC
  - UFR Lettres, Langues et Sciences Humaines : LLSH
  - UFR Psychologie, Sciences sociales, Sciences de l'éducation : Psycho
  - UFR Sciences et Techniques des Activités Physiques et Sportives : STAPS
  - **UFR Sciences et Technologies : ST**

- 5 écoles et instituts :
  - **Institut Supérieur d'Informatique, de Modélisation et de leurs Applications : ISIMA**
  - Polytech' Clermont-Ferrand : CUST
  - Institut Universitaire de Formation des Maîtres d'Auvergne : IUFM
  - **Institut Universitaire de Technologie d'Allier : IUT de Montluçon**
  - Observatoire de Physique du Globe de Clermont-Ferrand : OPGC

UFR Sciences et Technologies

Département Mathématiques Informatique

- Formations :
  - Mathématiques : L1 à M2, Doctorat
  - Informatique : environ 450 étudiants en L et M
    - Licence Informatique (L1 à L3)
    - Licence Professionnelle MOIE (L3)
    - Master Informatique et Systèmes (M1 à M2)
      - Recherche (M2)
      - Professionnelle
        - Informatique Décisionnelle et Logicielle (M2)
        - Internet et Pilotage de Projets (M1 et M2)
    - Doctorat

Laboratoire d’Informatique, de Modélisation et d’Optimisation des Systèmes (LIMOS)

Scientific Identity :

Models and **Tools** for

- Design, Representation, Evaluation, Forecasting, Control, and Optimization

of

Complex Organizational Systems

- Transport, Telecoms, Manufacturing, Ecosystems, Biosystems

Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)

Multi-themes, structured by axis:

- Axis 1 : Models and Algorithms of Decision Support-Systems
- Axis 2 : **Information and Communication Systems**
  - Theme 1 : Sensor Networks
  - Theme 2 : **Data, Services and Interoperability**
- Axis 3 : Manufacturing Systems

Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)

Theme : Data, Services and Interoperability

- Two main problems :
  - ➡ Interoperability (data and program)
    - ✦ Modeling and management of web services
    - ✦ Data management in distributed environment
    - ✦ Data Modeling and Integration
  - ➡ Knowledge Extraction
    - ✦ **Semantics in knowledge extraction**
    - ✦ Concise Representation of Knowledge
    - ✦ Usage traces, Bioinformatics

Laboratoire d’Informatique, de Modélisation et d’Optimisation des Systèmes (LIMOS)

Theme DSI --- some Projects :

LabEx IMobS^3 2012-2022

- "Innovative Mobility: Smart and Sustainable Solutions" --- Track #2

Investissement d’avenir BreedWheat 2011-2020

- Wheat, Biology, Bioinformatics, Genetics, Genomics, Ecophysiology, high throughput phenotyping and genotyping, …

CNRS Mastodons PETASKY 2012-xxxx

- Management and exploration of massive scientific data from astronomy observations

Laboratoire d’Informatique, de Modélisation et d’Optimisation des Systèmes (LIMOS)

Theme DSI --- Collaborations :

UQAM

- Roger Nkambou --- (2006-)
  - Educational Data Mining
  - PhD of Philippe Fournier-Viger
    - SPMF : an open-source data mining library
  - ….

- Abdoulaye Baniré Diallo
  - Bioinformatics
  - IJCAI 2015 workshop on Bioinformatics
  - …

Outline

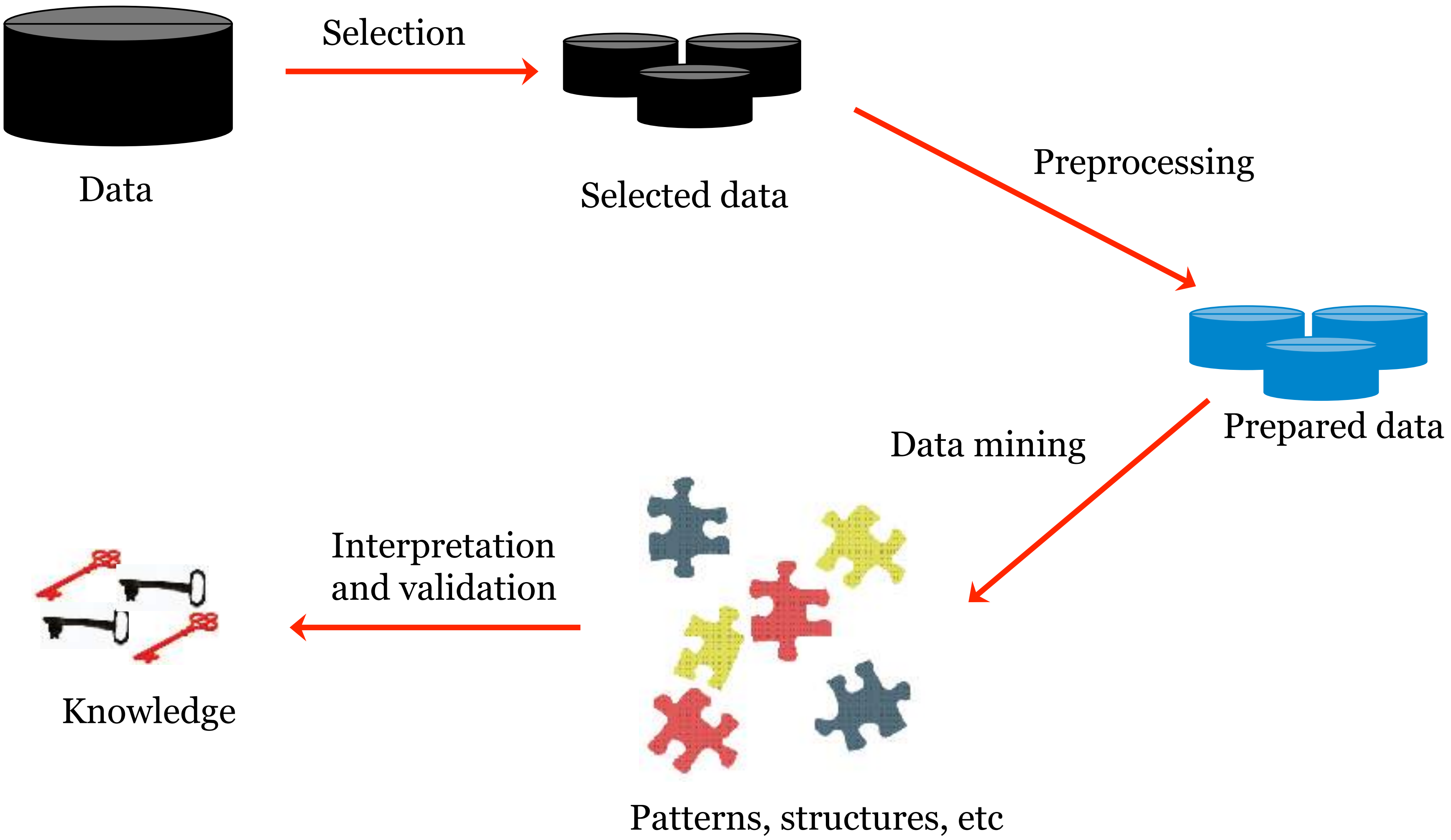
1. Pattern Mining
2. Pattern mining and Preferences
3. Skyline association rules
4. Discussion

Pattern Mining

"The real power of human thinking is based on recognizing patterns"

Ray Kurzweil

Pattern Mining : KDD process



Pattern Mining

Goal : Find all patterns that adhere to some constraint (Mannila & Toivonen, 1997).

Formally :

Given a **pattern language L**, a **constraint p** on **data D** (p can be a conjunction of constraints), find the set of all **patterns π** in the language L that satisfy the constraint p, that is, the theory :

$$Th(L, p, D) = \{\pi \in L \mid p(\pi, D) \text{ is true}\}.$$

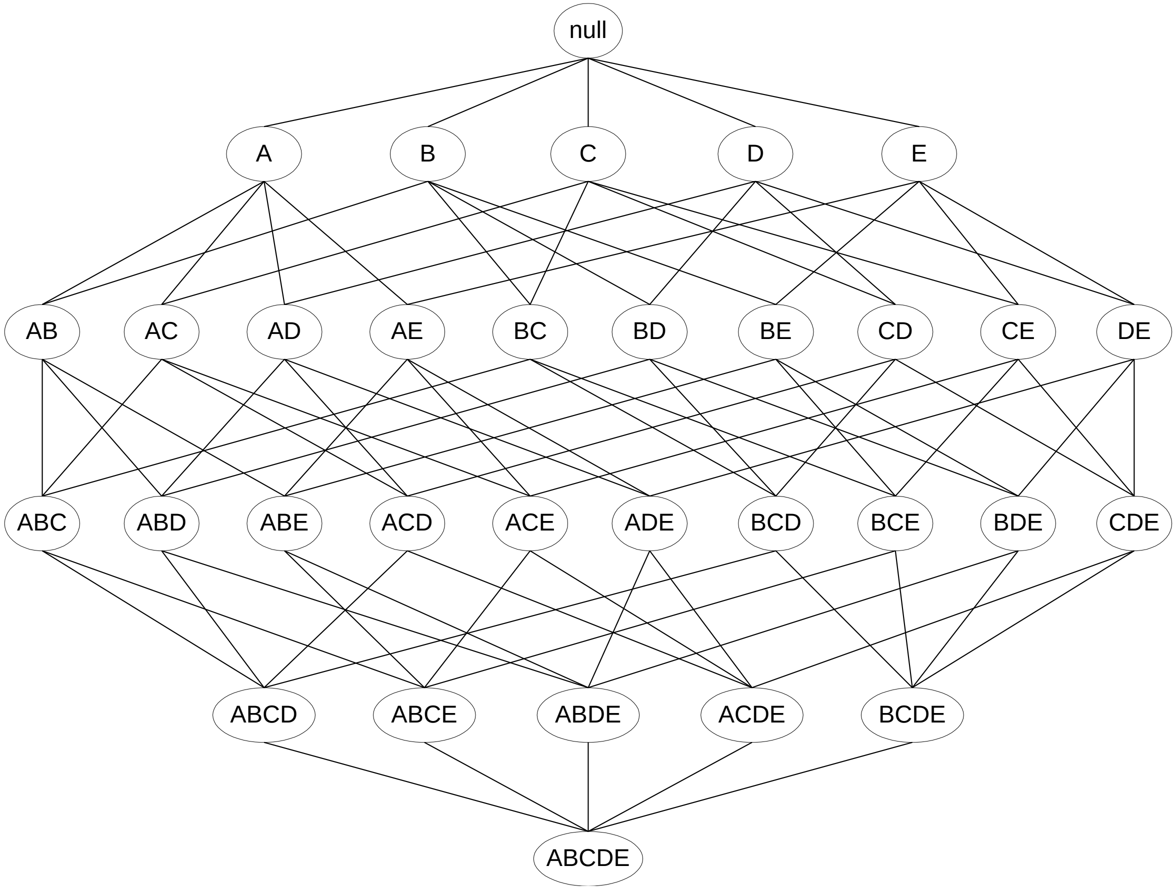
Example : Finding **frequent itemsets** (Agrawal & al. 1993)

- Data : Set of transactions, each transaction contains a set of items
- Pattern Language : Space of all itemsets
- Constraint : Frequency(itemset) >= Minimum Support threshold

Pattern Mining

Patterns: (Association Rules), Sequences, Trees, Graphs, ...

Search principle : Anti-monotonicity property



Pattern Mining

Problems with Frequency :

- Too many patterns, Not restrictive enough to find interesting patterns

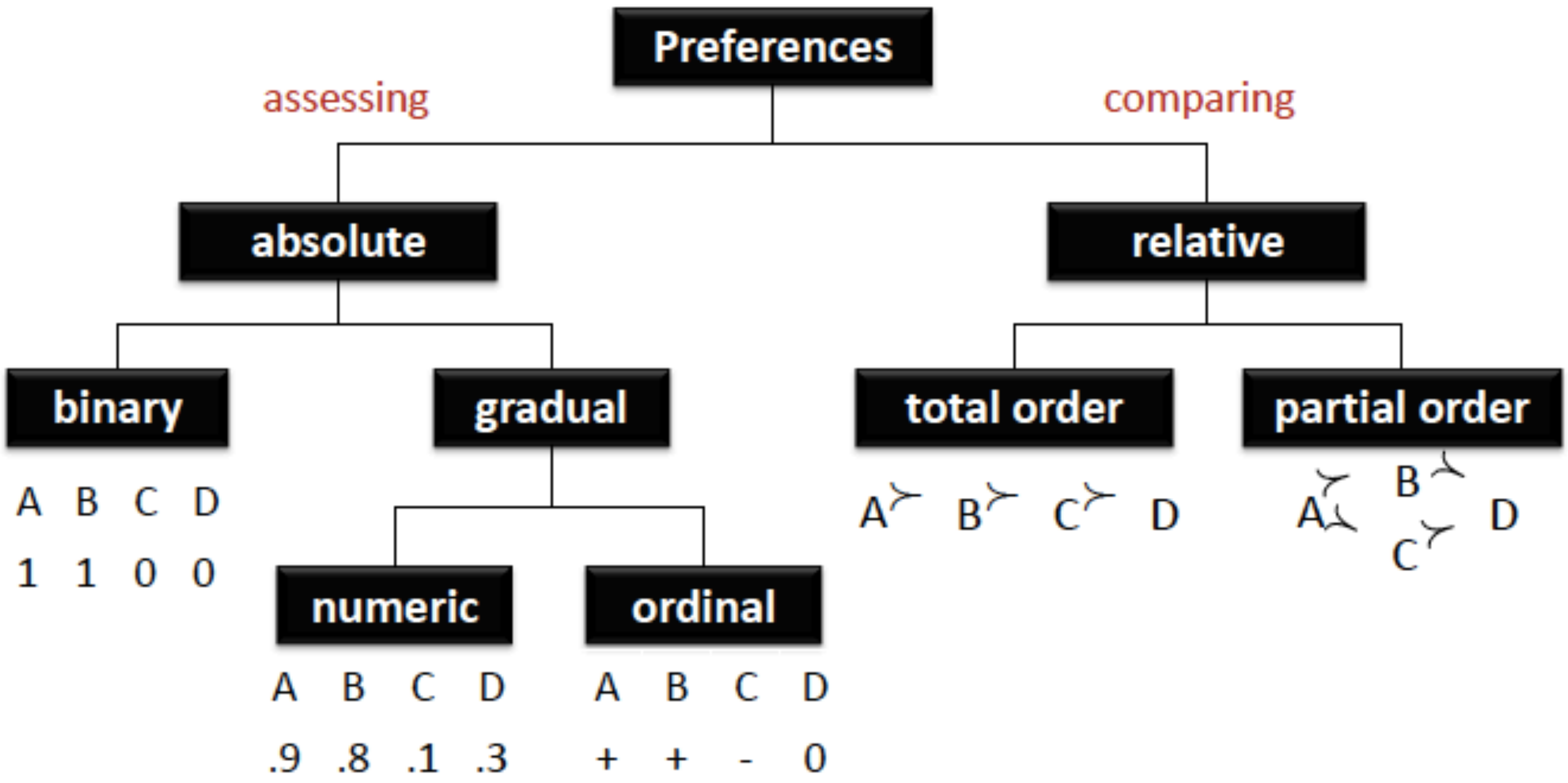
One solution : Additional types of constraints

- Anti-monotonicity framework (Bayardo etal., 2000) :
  - Maximum size of an itemset
- Condensed representations :
  - Maximal itemset (Bayardo, 1998), Closed itemset (Pasquier etal., 1999), Free itemsets (Boulicaut etal., 2000), ...
- Discriminative patterns
  - Find itemsets that discriminate between the two databases.
  - Also known as : correlated itemset mining, interesting itemset mining, contrast set mining, emerging itemset mining, subgroup discovery.

Pattern Mining and Preferences

Preferences : A set of alternatives

J. Fürnkranz & E. Hüllermeier, ECAI'12 Tutorial



Pattern Mining and Preferences

Preferences : J. Fürnkranz & E. Hüllermeier, ECAI'12 Tutorial

- binary vs. graded (e.g., relevance judgments vs. ratings)
- absolute vs. relative (e.g., assessing single alternatives vs. comparing pairs)
- **explicit vs. implicit** (e.g., direct feedback vs. click-through data)
- structured vs. unstructured (e.g., ratings on a given scale vs. free text)
- **single user vs. multiple users** (e.g., document keywords vs. social tagging)
- …

Objective :

- A group of agents expresses preferences, and comes to a common decision

Where :

- Social Sciences (voting and choice theory); Economics and decision theory; Multiple criteria and decision making; Operations research;
- Recommender Systems; Adaptative user interfaces; …

Pattern Mining and Preferences

Preferences :

- Borda (1781), Condorcet (1785)
- K. J. Arrow (1963), A. K. Sen (1976)

- Rossi, F., Venable, K., Walsh, T.: 2011, A short introduction to preferences: Between artificial intelligence and social choice

- http://www.preference-learning.org/

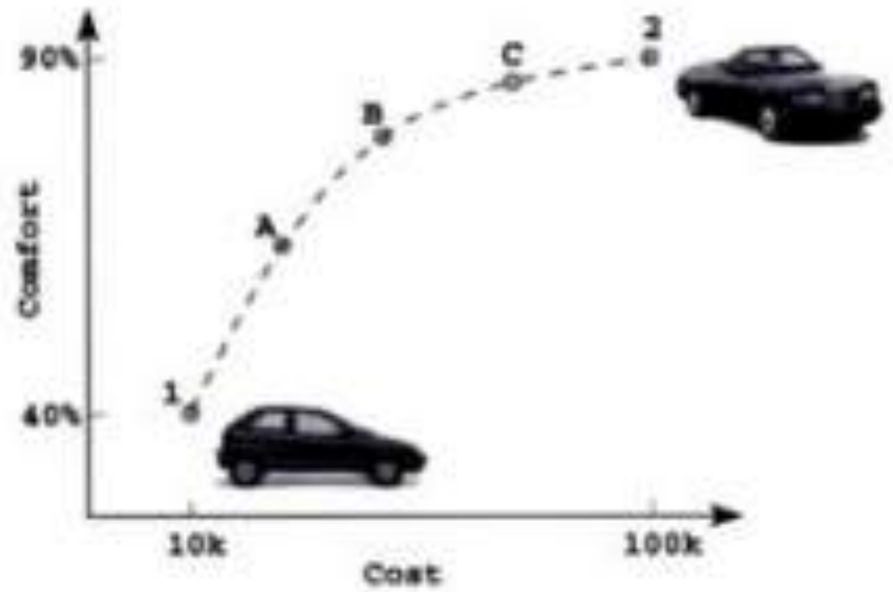
From individual preferences to collective preferences :

- Main operation : Aggregation
  - Non-dominance

Pattern Mining and Preferences

Preferences and non dominance :

- Skyline
- Find good compromises (or trade-offs) instead of a single solution (global optimization).
- Pareto frontier



Readings :

- Kung, H. T., Luccio, F., and Preparata, F. P. 1975. On Finding the Maxima of a Set of Vectors. J. ACM 22, 4 (Oct. 1975), 469-476

Pattern Mining and Preferences

Skyline and Databases :

- Skyline queries
    - S. Börzsönyi *et al.*, (2001) 'The skyline operator', in ICDE.
    - …

Skyline and Data Mining :

- SkyCubes
    - J. Pei *et al.* (2005), 'Catching the best views of skyline: A semantic approach based on decisive subspaces', in VLDB.
    - W. Ugarte et al. (2014), 'Computing SkyCubes Patterns', in ECAI.
- Skyline patterns : mainly itemsets
    - A. Soulet *et al.* (2011), 'Mining dominant patterns in the sky', in IEEE ICDM.

Association rules

- **Objective :**

  Discovery of useful correlations between attributes (items) of a database.

- **Applications :**
  - Market basket analysis
  - Intrusion detection
  - Risk Management

Association rules

[Agrawal and Srikant, 1993]

- Mining all association rules X $\longrightarrow$ Y, X and Y being disjoint itemsets, such that:
  - $Supp(X \longrightarrow Y) = P(X \cup Y) \geq \sigma_s$
  - $Conf(X \longrightarrow Y) = P(Y/X) \geq \sigma_c$

ID	Items
1	a,c,d
2	b,c,e
3	a,b,c,e
4	b,e
5	a,b,c,e

Transaction Database

$\sigma_s = 2/5$ et $\sigma_c = 1/2$

$a \rightarrow bce$ [2/5;2/3]
$b \rightarrow ace$ [2/5;1/2]
$c \rightarrow abe$ [2/5;1/2]
$e \rightarrow abc$ [2/5;1/2]
$ab \rightarrow ce$ [2/5;1]
$abc \rightarrow e$ [2/5;1]
...

Association Rules

- Approaches using supp/conf generate a large number of rules especially for large databases (mining irrelevant rules, etc).

- The expert is unable to :
  - determine the most interesting association rules.
  - ….

→ Need for an efficient evaluation of rules to select those that are actually relevant.

Association Rules

Datasets	**minsup**	**exact Rules**	**exact Gen. Base**
T10I4D100K	0,5 %	0	**0**
Mushrooms	30 %	7 476	**543**
C73D10K	90 %	52 035	**1 369**

Datasets	**minsup**	**minconf**	**Approx. Rules**	**Approx. Gen. Base**
T10I4D100K	0,5 %	70 % 50 %	20 419 21 686	**4 004** **4 191**
Mushrooms	30 %	70 % 50 %	37 671 56 703	**1 221** **1 481**
C73D10K	90 %	95 % 85 %	1 606 726 2 053 936	**5 680** **5 718**

Association Rules

- Many approaches:
  - Using measures :
    - objective interestingness measures [Freitas, 1999].
    - subjective interestingness measures [Silberschatz & Tuzhilin, 1995].
  - Redundancy analysis [Lehn et al., 1998].
  - Visualization tools [Blanchard et al., 2003].

Association Rules

- **Objective interestingness measures :**

  Numerically quantify the interest of a rule r:

$$\mathcal{R}: \quad m \mapsto \mathbb{R}$$

$$r \mapsto r[m]$$

- **Utility :**
  - Ranking rules.
  - Filtering rules using a threshold.

Association Rules

- Some objective measures :

SUP	p_{ab}	KAPPA	$2\frac{p_{ab}-p_a p_b}{p_a p_{\bar{b}}+p_{\bar{a}} p_b}$
CONF	$\frac{p_{ab}}{p_a}$	LIFT	$\frac{p_{ab}}{p_a p_b}$
R	$\frac{p_{ab}-p_a p_b}{\sqrt{p_a p_b p_{\bar{a}} p_{\bar{b}}}}$	FUKUDA	$np_a(\frac{p_{ab}}{pa}-\sigma_c)$
LOE	$1-\frac{p_{a\bar{b}}}{p_a p_{\bar{b}}}$	JAC	$\frac{p_{ab}}{p_{a\bar{b}}+p_b}$
CENCONF	$\frac{p_{ab}-p_a \times p_b}{p_a}$	SEB	$\frac{p_{ab}}{p_{a\bar{b}}}$
CONV	$\frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$	PS	$n(p_{ab}-p_a p_b)$
LAP	$\frac{p_{ab}+1/n}{p_a+2/n}$	LC	$\frac{p_{ab}-p_{a\bar{b}}}{p_b}$
FB	$\frac{p_{ab}/p_b}{p_{a\bar{b}}/p_{\bar{b}}}$	ECR	$\frac{p_{ab}-p_{a\bar{b}}}{p_{ab}}$
IG	$\log(\frac{p_{ab}}{p_a p_b})$	ZHANG	$\frac{p_{ab}-p_a p_b}{\max\{p_{ab} p_{\bar{b}}, p_b p_{a\bar{b}}\}}$
GAN	$2(p_{ab}/p_a-0.5)$	.....	.....

Association Rules

- Problems caused by the abundance of measures : ≈ 60.

- **Heterogeneity :** the outputs of evaluations vary from a measure to another one.

  → the output rules are not the same.

- Appearance of several approaches that adopt the hypothesis:

  Selecting the “right” rules requires using the “right” measures

  [Hilderman and Hamilton, 2000, Tan et al., 2002, Lenca et al., 2002].

- Characterizing measures according to properties:
  - No symmetry,
  - Tolerance to counter-examples,
  - Value at independence,
  - etc

Association Rules

Disadvantages of using properties

- Translation of preferences of expert on properties is a non-trivial task.
- The problem of heterogeneity of evaluation persists especially in the case of selecting multiple "good" measures.
- Difficulty for the user to specify **thresholds values** for the measures.

Association Rules

- Why favor one measure from another?
- Find a trade-off between the measure results.

- Discover interesting association rules:
  - Without favoring or excluding any measure.
  - By avoiding the problem of the threshold value specification.

- Using dominance relationship.

Association Rules : Dominance

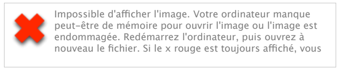
Principle

- *A rule r is dominated by another one r', if for all measures, r is less relevant than r'.*

- *The rule r is discarded from the result, not because it is not relevant for one of the measure but because it is not interesting according to the combination of all measures.*

Association Rules : Dominance

	$m1$	$m2$	$m3$	$m4$
$r1$	0,7	0,8	0,5	0,4
$r2$	0,8	0,7	0,6	0,5
$r3$	0,6	0,6	0,4	0,3
$r4$	0,6	0,4	0,4	0,2
$r5$	0,4	0,5	0,3	0,1
$r6$	0,3	0,4	0,4	0,3
$r7$	0,5	0,8	0,6	0,4
$r8$	0,5	0,5	0,4	0,3

Example : 

- **Dominance relationship:** Let $M = \{m_1, m_2, \ldots, m_k\}$ the set of measures. Given two rules r and r', the dominance relationship according to the set of measures M is defined as follows:

$$\textit{iff} \quad r[m_1] \geq r'[m_1] \;\; and \;\; r[m_2] \geq r'[m_2] \;\; and \;\; \ldots \; and \;\; r[m_k] \geq r'[m_k]$$

When 

and $\exists\, m_i \in M$ such that $r[m_i] > r'[m_i]$, the dominance is strict

Association Rules : Dominance

- The strict dominance is irreflexive : $r \not\succ r$

- The dominance is transitive: $\forall\, r, r', r'' \in \mathcal{R}$ if $r \succ r'$ and $r' \succ r''$ then $r \succ r' \succ r''$

- If a rule r dominates another one r', then r is equivalent or better than r' for all measures. The dominated rules (at least one) are not relevant and have to be eliminated from the operator SkyRules+.

- **SkyRules$^+$:** Let $\mathcal{R}$ the set of association rules. the SkyRules$^+$ of $\mathcal{R}$ according to $\mathcal{M}$ is the set of undominated rules of $\mathcal{R}$ with respect to $\mathcal{M}$:

$$\mathrm{Sky}^{+}_{\mathcal{M}}(\mathcal{R}) = \{r \in \mathcal{R} \;/\; \nexists\, r' \in \mathcal{R} \backslash r,\; r' \succ r\}$$

Association Rules : Dominance

	$m1$	$m2$	$m3$	$m4$
$r1$	0,7	0,8	0,5	0,4
$r2$	0,8	0,7	0,6	0,5
$r3$	0,6	0,6	0,4	0,3
$r4$	0,6	0,4	0,4	0,2
$r5$	0,4	0,5	0,3	0,1
$r6$	0,3	0,4	0,4	0,3
$r7$	0,5	0,8	0,6	0,4
$r8$	0,5	0,5	0,4	0,3

$\text{Sky+}_{\mathcal{M}}(\mathcal{R}) = \{r1, r2, r7\}$

- A naïve approach consists on comparing all rules between them. The number of comparisons is equal to :

$$|M| \times \sum_{1}^{i=|R|} (i-1) = |M| \times \frac{|R|^2 - |R|}{2}$$

- Reduce the number of rules using generic bases, but this is generally not possible because some measures, are not condensable.

Association Rules : Dominance

	m1	*m2*	*m3*	*m4*	DegSim
r1	0,7	0,8	0,5	0,4	0,075
r2	0,8	0,7	0,6	0,5	0,025
r3	0,6	0,6	0,4	0,3	0,200
r4	0,6	0,4	0,4	0,2	0,275
r5	0,4	0,5	0,3	0,1	0,350
r6	0,3	0,4	0,4	0,3	0,325
r7	0,5	0,8	0,6	0,4	0,100
r8	0,5	0,5	0,4	0,3	0,250

Example:
$r^{\top}$ = < 0.8, 0.8, 0.6, 0.5 >

- **Reference association rule** $r^{\top}$ is a fictitious rule that dominates all the rules of $\mathcal{R}$.

- **Degree of similarity :** Given two rules $r, r' \in \mathcal{R}$, the degree of similarity between r and r' with respect to a set of measures M is defined as follows:

$$\text{DegSim}(r, r') = \frac{\sum_{m \in M} \| r[m] - r'[m] \|}{|M|}$$

- **Proposition :** Let $r^* \in \mathcal{R}$, be a rule having the minimal degree of similarity with respect to $r^{\top}$, then $r^* \in \text{Sky+}_M(\mathcal{R})$.

Association Rules : Dominance

	$m1$	$m2$	$m3$	$m4$	DegSim
$r1$	0,7	0,8	0,5	0,4	0,075
$r2$	0,8	0,7	0,6	0,5	0,025
$r3$	0,6	0,6	0,4	0,3	0,200
$r4$	0,6	0,4	0,4	0,2	0,275
$r5$	0,4	0,5	0,3	0,1	0,350
$r6$	0,3	0,4	0,4	0,3	0,325
$r7$	0,5	0,8	0,6	0,4	0,100
$r8$	0,5	0,5	0,4	0,3	0,250

Example : $r^* = r2$

$S_1 = \varnothing$

$S_2 = \{r1, r7\}$,

$S_3 = \{r7\}$,

$S_4 = \varnothing$

After identifying r^* :

1- The rules dominated by r^* must be discarded

2- Decompose the remaining rules into (k-1) sets $S_1, S_2, \ldots, S_{k-1}$ such that

$$S_i = \{r \in \mathcal{R} \mid r^* \nsucc r \text{ et } r[m_i] \geq r^*[m_i]\}$$

- **Proposition :** Let r an undominated rule. if $r \notin S_i$ then $\forall\, r' \in S_i, r \nsucc r'$

Association Rules : Experiments & Results

UNDOMINATED RULES *vs* TB RULES AND ALL RULES

Datasets (min_{freq} %)		{Conf;Loev}	{Conf;Pearl}	{Conf;Recall}	{Conf;Zhang}	{Conf;Pearl;Recall}	{Conf;Loev;Zhang}	{Conf;Loev;Pearl; Recall;Zhang}
Diabetes (10,00)	Sky-R	3411	9	6651	2996	9	171	171
	TB-R	59314	58124	59206	59309	44813	44602	42126
	A-R	62132	62132	62132	62132	62132	62132	62132
Flare (10,00)	Sky-R	4975	48	4978	4857	48	48	48
	TB-R	56163	57101	56451	54524	53197	53116	52819
	A-R	57476	57476	57476	57476	57476	57476	57476
Iris (0,00)	Sky-R	246	246	246	246	246	246	246
	TB-R	440	440	440	440	440	440	440
	A-R	440	440	440	440	440	440	440
Monks1 (1,00)	Sky-R	768	1	788	656	1	1	1
	TB-R	60417	60692	59418	59452	58904	58811	58327
	A-R	62184	62184	62184	62184	62184	62184	62184
Monks2 (1,00)	Sky-R	279	3	215	202	3	3	3
	TB-R	59611	59702	59568	59544	59103	58917	58662
	A-R	59976	59976	59976	59976	59976	59976	59976
Monks3 (1,00)	Sky-R	1028	2	713	781	4	2	2
	TB-R	58662	58369	57922	58436	57816	57734	56038
	A-R	59304	59304	59304	59304	59304	59304	59304
Nursery (2,00)	Sky-R	497	2	304	342	8	2	2
	TB-R	23872	23901	23875	23417	23176	22806	22139
	A-R	25062	25062	25062	25062	25062	25062	25062

Association Rules : Experiments & Results

GAIN OF THE UNDOMINATED RULES

Measures	Average number of Sky-R	Average number of TB-R	Average gain of Sky-R
{Conf;Loev}	2623,50	48308,75	18,41
{Conf;Pearl}	43,37	40908,12	943,23
{Conf;Recall}	2913,75	48094,00	16,50
{Conf;Zhang}	2399,00	47658,50	19,86
{Conf;Loev;Recall}	43,37	45347,12	1045,58
{Conf;Pearl;Zhang}	63,62	45192,75	710,35
{Conf;Loev;Pearl;Recall;Zhang}	63,62	44309,62	696,47

Association Rules : Top-K Dominance

- Undominated rules might not be enough answer to a personalized user query.

- The user often needs a specified number K of relevant rules, such that two cases are to be distinguished:

  - K > number of undominated rules

    - adding dominated rules

  - K < number of undominated rules

    - Selecting rules from the set of undominated ones

→ A ranking process must be performed on the whole set of rules (dominated or undominated).

Association Rules : Top-K Dominance

- The ranking process must satisfy the two following objective conditions:

  1. *Any dominated rules cannot be better ranked than an undominated one.*

  2. *Two undominated rules must be ranked based on degree of similarity with respect to reference rule.*

Association Rules

Ranking association rules

- **Succession relationship:** let two rules $r, r' \in \mathcal{R}$, r succeeds r', denoted by

  $r \lhd r'$, iff $r' \succ r$ and $\nexists r'' \in \mathcal{R}$ such that $r' \succ r'' \succ r$

- **Succession operator:** Let E be a set of rules such that $E \subseteq \mathcal{R}$. The successor of E in $\mathcal{R}$ with respect to $\mathcal{M}$, is defined as follows:

  $\text{Succ}_{\mathcal{M}}(E,\mathcal{R}) = \{r \in \mathcal{R} \setminus E \mid \exists\, r' \in E,\ r \lhd r' \wedge \nexists r'' \in E,\ (r'' \succ r \wedge r \ntriangleleft r'')\ \}$

- **Proposition :** Given a set of measures $\mathcal{M}$, for all $E \subseteq \mathcal{R}$ the following relation is fulfilled :

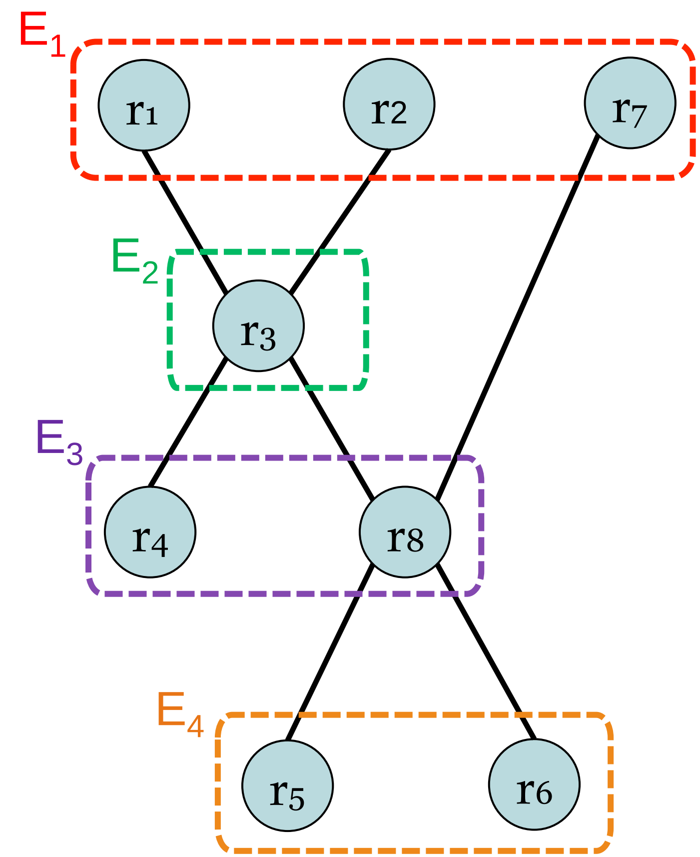
$$\text{Succ}_{\mathcal{M}}(\text{Sky}^{+}_{\mathcal{M}}(E),E) = \text{Sky}^{+}_{\mathcal{M}}(E \setminus \text{Sky}^{+}_{\mathcal{M}}(E))$$

Association Rules : Top-K Dominance

- The relation of succession reflects a total preorder relation between association rules

- First compute the set E of better rules that are not dominated by any other rule.
  Then compute the successor set of E.

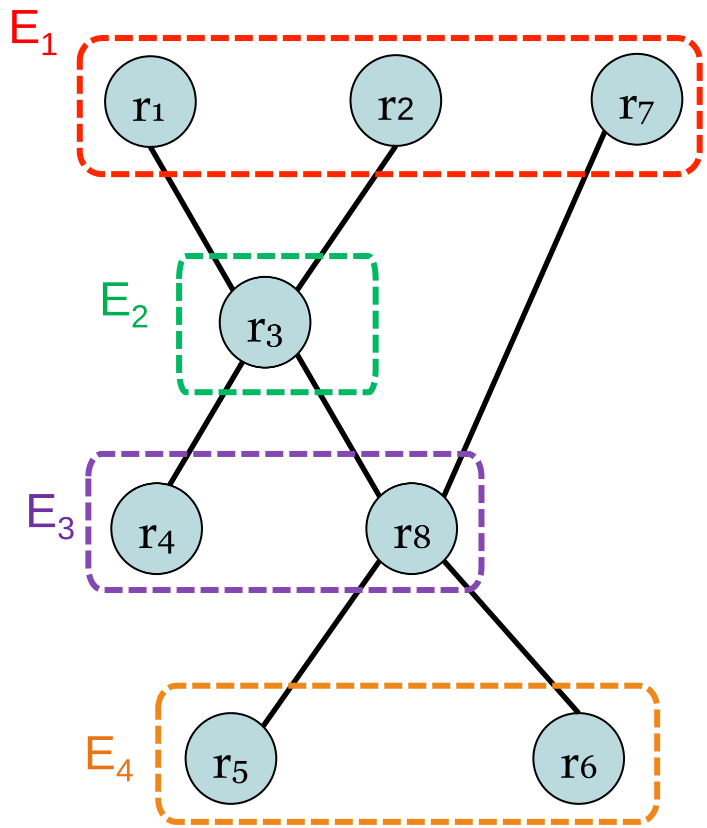
Input : The set of rules $\mathcal{R}$ & the set of measures $\mathcal{M}$
Output : A complete preorder over $\mathcal{R} \times \mathcal{R}$
Begin
p $\leftarrow$ 0;
    While $\mathcal{R} \neq \emptyset$ **do**
        p $\leftarrow$ p+1;
        $E_p \leftarrow Sky+_{\mathcal{M}}(\mathcal{R})$;
        $\mathcal{R} \leftarrow \mathcal{R} \setminus E_p$
    End While
Return $(E_1, \ldots, E_p)$
End

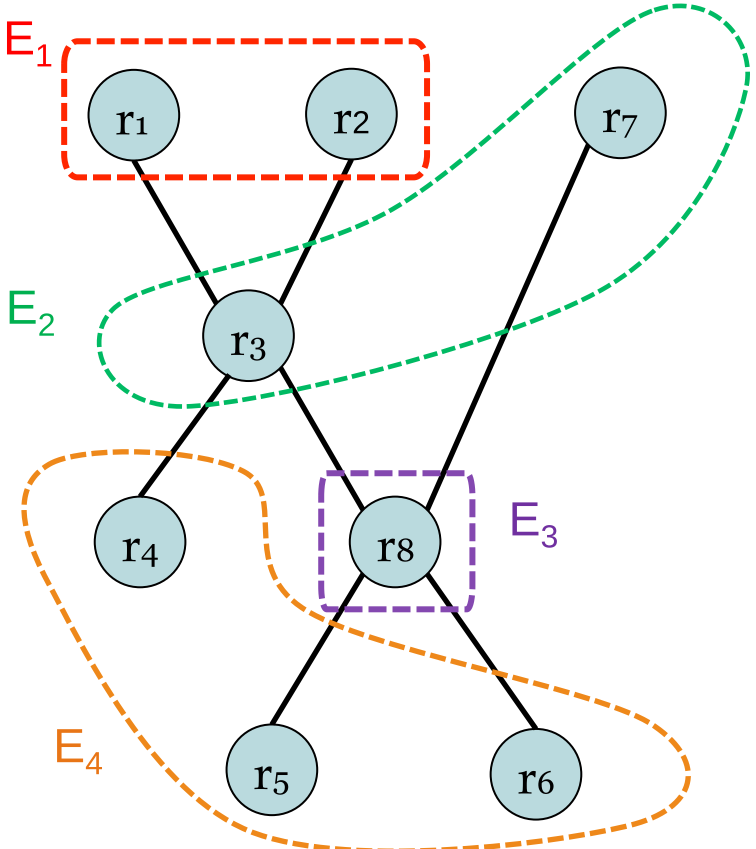


Association Rules : Top-K Dominance

- **SkyRules$^-$:** Let $\mathcal{R}$ be the set of association rules. the SkyRules$^-$ of $\mathcal{R}$ according to $\mathcal{M}$ is the set of undominated rules of $\mathcal{R}$ with respect to $\mathcal{M}$ which don't dominates any other rule, following the set $\mathcal{M}$:

$$Sky^-_{\mathcal{M}}(\mathcal{R}) = \{r \in \mathcal{R} \mid \nexists\ r' \in \mathcal{R} \backslash r,\ r \succ r'\}$$





Conclusion

[Stephen H. Muggleton, 22 Mars 2006]
Nature (www.nature.com)
'2020 Computing: Exceeding human limits'

« Scientists are turning to automated processes and technologies in a bid to cope with ever **higher volumes of data**

... It is clear that the **future of science involves** the expansion of automation in all its aspects: *data collection*, *storage of information*, **hypothesis formation** and *experimentation*

... However, to reap the full benefits it **is essential** that developments in high-speed automation are not introduced at the expense of **human understanding** and insight. ... »

Conclusion

- **First problem : huge number of rules**

  - The use of interestingness measures

- **Second problem : abundance of measures**

  - selection process of association rules for all measures
  - ranking process of association rules for all measures

  } Using Dominance relationship

Further work

- Setting up an approach aiming at discovering undominated rules during the phase of the extraction rules which will improve the performance of the SKYRULE algorithm.

- Take into account similarity between rules in order to eliminate redundant undominated rules

	m1	m2	
r1 : a b → e	*0.7*	*0.8*	
r2 : a → e	*0.6*	*0.5*	✗
r3 : c → d	*0.6*	*0.5*	✓

International Journal on Artificial Intelligence Tools
Vol. 23, No. 4 (2014) 1460011 (22 pages)

DOI: 10.1142/S0218213014600112

World Scientific
www.worldscientific.com

Mining Undominated Association Rules Through Interestingness Measures

Slim Bouker etal.

Clermont Université, Université Blaise Pascal
LIMOS, BP 10448, F-63000 Clermont-Ferrand, France
CNRS, UMR 6158, LIMOS, F-63173 Aubière, France
University of Tunis El Manar, Faculty of Sciences of Tunis
LIPAH, 2092, Tunis, Tunisia

Best student paper award
at ICTAI'2012 conference (Athens, Greece)

Supported by :
- PHC French-Tunisian project CMCU-Utique EXQUI,
- Auvergne Region

THANKS

IEEE COMPUTER SOCIETY

Biological & Artificial Intelligence FOUNDATION

INTERNATIONAL CONFERENCE on TOOLS with ARTIFICIAL INTELLIGENCE (ICTAI-12)

November 07-09, 2012,
Athens, Greece

Present to

Slim Bouker, Rabi Saidi, Sadok Ben Yahia and Engelbert Mephu Nguifo

the **ICTAI-12 Best Student Paper Award** for the paper "**Ranking and selecting association rules based on dominance relationship**"

IEEE Computer Society, BAIF and ICTAI Awards Committee

Algorithm

Algorithm 1: SKYRULE

Input: $\Omega = (\mathcal{R}, \mathcal{M})$

Output: Sky: set of undominated rules of Ω.

Begin
- $Sky \leftarrow \emptyset$
- $C \leftarrow \mathcal{R}$
- $\mathcal{E} \leftarrow \{\mathcal{R}\}$
- **While** $C \neq \emptyset$ **do**
  - $r^* \leftarrow r \in C$ **having** $min(DegSim(r,r^{\perp}))$
  - $C \leftarrow C \setminus \{r^*\}$
  - **for** i=1 *to* k **do**
    - $s_i^{r^*} \leftarrow \emptyset$
  - $Sky \leftarrow Sky \cup \{r^*\}$
  - **Foreach** $e \in \mathcal{E}$ *such that* $r^* \in s$ **do**
    - **Foreach** $r \in e$ **do**
      - **If** $r^* \succ r$ **then**
        - $C \leftarrow C \setminus \{r\}$
      - **Else**
        - **for** i=1 *to* k **do**
          - **If** $r[m_i] \succ r^*[m_i]$ **then**
            - $s_i^{r^*} \leftarrow s_i^{r^*} \cup \{r\}$
    - $\mathcal{E} \leftarrow \mathcal{E} \setminus \{e\}$
  - $\mathcal{E} \leftarrow \mathcal{E} \cup \{s_1^{r^*}, \ldots, s_k^{r^*}\}$
- **return** Sky

End