Constraint-based mining: preliminary results on exploiting expert models

Jean-Francois Boulicaut
LIRIS UMR 5205 Team DM2L, INSA de Lyon, France

(Joint work with colleagues from PPME, University of New Caledonia: Frédéric Flouvat, Nazha Selmaoui-Folcher, and Jérémy Sanhes)

Dagstuhl Seminar 14411, Wadern (D), October 7th 2014

Context: From Data to Knowledge by means of Patterns

- Supporting KDD processes in real-life settings
  - Data Mining: pattern discovery from (more or less) big data
  - Supporting the whole knowledge discovery process?
    ... not only designing efficient pattern discovery algorithms
  - Humans in the loop: a querying vision on KDD processes
- Data Science projects
  - Biology, Logistics, Solar Energy (Physics), Sociology
  - Which kind of scientific cooperation? Is it possible to support solution co-design?

Our main directions of research

• Formalizing as many steps as possible of KDD processes within the constraint-based data mining framework, i.e., designing pattern domains and implementing solvers, is a needed step towards inductive databases

$$\{\phi \in \mathcal{L} \mid \mathcal{C}_1(\phi, \mathcal{R}) \wedge \ldots \wedge \mathcal{C}_n(\phi, \mathcal{R})\}$$

• Defining the data $\mathcal{R}$, the pattern language $\mathcal{L}$, the primitive constraints $\mathcal{C}_i$, and the way to combine them, ...

• A declarative view (specifying a priori relevancy)

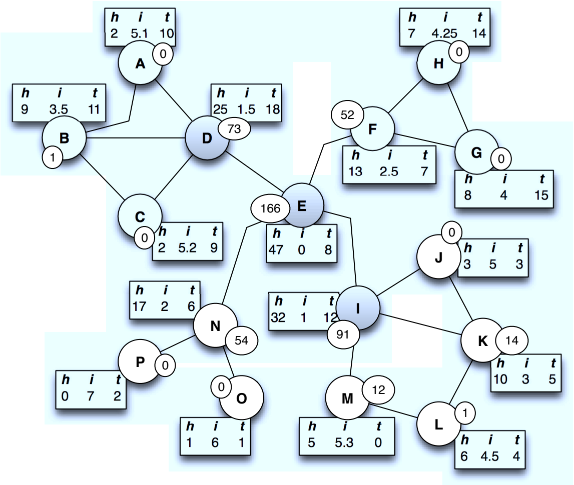
• A computational view (computing patterns)

Examples of pattern domains and inductive queries

$$\{\phi \in \mathcal{L} \mid \mathcal{C}_1(\phi, \mathcal{R}) \wedge \ldots \wedge \mathcal{C}_n(\phi, \mathcal{R})\}$$

- Examples
  - Constrained clustering of objects given sets of features
  - Detecting dense subgraphs or topological patterns in an (attributed) graph
  - Discovering patterns in Boolean data like formal concepts or their generalizations
- Problems
  - Ensuring pattern relevancy and subjective interestingness
  - Tractable evaluation in practice for more or less generic solvers
  - ...

An example for topological patterns



Co-authorship network.

$\{h^{+}, i^{-}, \textit{betweness}^{+}\}$

Prado et al. IEEE TKDE, 2013.

The example of closed n-sets

Let $(\mathcal{D}^i)_{i=1..n}$ n finite sets and $\mathcal{R} \subseteq \times_{i=1..n}\mathcal{D}^i$ an n-ary relation.

Definition

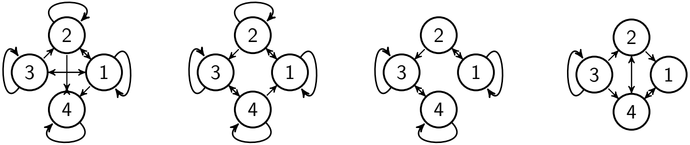
$\forall (X^1, \ldots, X^n) \subseteq \times_{i=1..n}, \mathcal{D}^i, (X^1, \ldots, X^n)$ is a closed n-set if and only if:

- $\mathcal{C}_{\text{connected}}(X^1, \ldots, X^n) \equiv \times_{i=1..n} X^i \subseteq \mathcal{R}$
- $\mathcal{C}_{\text{closed}}(X^1, \ldots, X^n) \equiv \forall i = 1..n, \forall x \in \mathcal{D}^i \setminus X^i,$ $\neg \mathcal{C}_{\text{connected}}(X^1, \ldots, X^i \cup \{x\}, \ldots, X^n)$

$$\{\phi \in \times_{i=1..n} 2^{\mathcal{D}^i} \mid \mathcal{C}_{\text{connected}}(\phi, \mathcal{R}) \wedge \mathcal{C}_{\text{closed}}(\phi, \mathcal{R}) \wedge \ldots\}$$

E.g., $(\{c_1, c_2\}, \{p_1, p_2\}, \{s_1, s_2\})$

Illustration when encoding relational graphs



	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
d_1	1	1	1	1	1	1			1	1						1
d_2	1	1		1	1	1	1		1		1		1			1
d_3	1	1	1				1	1			1	1		1	1	1
d_4				1	1		1	1			1	1	1	1		
	t_1				t_2				t_3				t_4			

- $(\{d_1, d_2\}, \{a_1, a_2\}, \{t_1, t_2\})$
- Is it relevant?
- It satisfies $C_{connected} \wedge C_{closed}$, enforcing other constraints?

Relevance constraints for dynamic graph mining

Symmetry constraint

A 3-set (N^1, N^2, T) is symmetric $\Leftrightarrow N^1 \subseteq N^2 \wedge N^2 \subseteq N^1$.

Maximal cross-graph cliques are extracted.

	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
d_1	1	1	1	1	1	1			1	1						1
d_2	1	1		1	1	1	1		1		1		1			1
d_3	1	1	1				1	1			1	1		1	1	1
d_4				1	1		1	1			1	1	1	1		
	t_1				t_2				t_3				t_4			

δ-contiguity constraint

Let $\delta \in \mathbb{R}_+$, a user-defined parameter. A 3-set (N^1, N^2, T) is δ-contiguous $\Leftrightarrow \forall t \in [\min(T), \max(T)], \exists t' \in T$ s.t. $|t - t'| \leq \delta$.

Cerf et al. Inductive databases and constraint-based data mining, 2010.

Pattern domains to solve data science problems

- Selecting well-studied pattern domains and over-exploiting experts for ad-hoc feature construction
  E.g., Encoding selected spatio-temporal properties in terms of boolean properties or in terms of an ad-hoc event alphabet when using itemsets or sequential patterns.
- Designing new pattern domains (e.g., for mining collections of trajectories)

Defining primitive constraints is the hard part of the job ;-)

Primitive constraints should enable to express (a) pattern semantics, (b) objective interestingness, and (c) subjective interestingness

Example of a fairly sophisticated inductive query

- Cross-graph preserved and unexpected clique mining
  - pattern semantics
  - objective interestingness
  - subjective interestingness

See also the tutorial by Cerf and Meira at ECML PKDD 2014

$$
\begin{aligned}
\{\phi \in \times_{i=1..n} 2^{\mathcal{D}^i} \mid \\
&\mathrm{C}_{\text{connected}}(\phi, \mathcal{R}) \wedge \mathcal{C}_{\text{closed}}(\phi, \mathcal{R}) \wedge \\
&\quad \mathrm{C}_{\delta\text{-contiguity}}(\phi, \mathcal{R}) \wedge \mathcal{C}_{\text{symmetry}}(\phi) \wedge \\
&\quad\quad \mathrm{C}_{\text{unexpected}}(\phi, \mathcal{R}) \ \wedge \ \ldots \}
\end{aligned}
$$

Humans in the loop?

- Writing (inductive) queries
  i.e., selecting data, selecting some primitive constraints and fixing their parameters, post-processing answers, often using domain and expert knowledge to remove uninteresting patterns
- Who? Computer scientists? Application domain experts?
  One step towards the co-design of data mining solutions is to exploit expert models
- A preliminary study for deriving constraints from expert models

$$F : dom(x_1) \times dom(x_2) \times \ldots \times dom(x_n) \to \mathbb{R}$$

$$\{\phi \in \mathcal{L} \mid \mathcal{C}_1(\phi, \mathcal{R}) \wedge \ldots \wedge \mathcal{C}_p(\phi, \mathcal{R}) \wedge \mathcal{C}_{p+1}(\phi, \mathcal{M}, \mathcal{R}) \wedge \ldots\}$$

Flouvat et al. ECAI 2014.

A case study about soil erosion understanding

Environmental problems in New Caledonia

- An exceptional environment: a biodiversity hotspot and a lagoon on the UNESCO World Heritage List
- Important mining projects (25% of the world's known nickel resources), a tropical climate with cyclones and bush fires

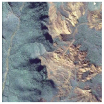
→ Strong soil erosion with an impact on ecosystems

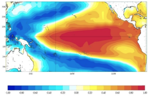
- Experts often express part of their knowledge into models e.g., to assess an erosion risk according to a set of environmental parameters
- Notice that only few parameters (among the collected data) are exploited

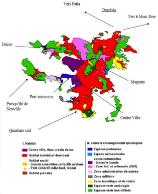
Related data science challenges

- Collecting Big Data
  - Satellite images, nature of the ground, populations, climate, vegetation, . . .

Zone	Date	Température	Humidité	Precipitations	Rafales
Z_1	25/01/2013	T_b	H_m	P_m	-
Z_1	26/01/2013	T_m	H_m	P_b	-
Z_1	27/01/2013	T_b	H_m	P_m	55
Z_2	25/01/2013	T_m	H_m	P_m	-
Z_2	26/01/2013	T_m	H_m	P_b	-
Z_2	27/01/2013	T_b	H_b	P_m	-
Z_3	25/01/2013	T_b	H_m	P_h	75
Z_3	26/01/2013	T_m	H_h	P_b	-
Z_3	27/01/2013	T_m	H_h	P_h	55





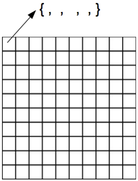


ANR FOSTER and CNRS MASTODONS Amadouer: computer scientists, geographs and geologists ;

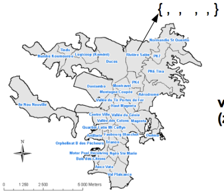
Our itemset mining context

- Data

raster (pixels)



et / ou



vecteur (zones)

Préparation et transformation des données (nettoyage, fusion, **discrétisation**, ...)

Attributs / **Objets**	Précipitation	Type de sol	Occupation du sol	...
objet 1	[0,1000]	latérite	mine	
objet 2		latérite	piste	
...				

- Patterns

Itemsets under constraints, e.g., {Annual rainfall=[0,1000], Soil type=laterite,Land cover=trail}

Models about soil erosion (1)

- Empirical models (e.g., USLE and Atherton)
  - Such models are often linear or polynomial ones
  - They are defined based on expert empirical knowledge and physical measures

Parameters	classes	values
Soil erodability x_{erod}	alluvium	1
	sand, laterite soil, ...	2
	swamp, nigrescent silt, ...	3
	ferruginous laterite soil,..	4
Soil land cover x_{occup}	water	0
	dense forest, wood production, ...	1
	sparse forest	2
	coconut plantation, non-forest area	3
	sugar cane farming	4

Parameters	classes	values
Slope (in %) x_{slope}	[0 , 3.5]	0.5
	[3.6 , 30]	1
	[31 , 50]	2
	[51 , 60]	3
	[60.1 , 100]	9
Rain intensity (in mm/year) x_{rain}	[0 , 2000]	1
	[2001 , 3200]	2
	[3201, 10 000]	3
Seasonality of rains (in mm) x_{season}	[0 , 70]	1
	[71, 200]	2

$$REP(x_{slope}, x_{rain}, x_{season}, x_{erod}, x_{occup}) = x_{slope} + x_{rain} + x_{season} + x_{erod} + x_{occup}$$

$$REP = \begin{cases} [6, 9.5[\longmapsto \text{LOW score} \\ [9.5, 11[\longmapsto \text{MEDIUM score} \\ [11, 12[\longmapsto \text{HIGH score} \end{cases}$$

Models about soil erosion (2)

- Physical models (e.g., WEPP and RMMF)
  - Often nonlinear and non-polynomial quantitative models based on physical properties

Parameters	domain of values
Soil detachment index (in g/J) x_K	depends on soil type
Annual rainfall (in mm) x_R	[0 , 12 000]
Proportion of rain stopped by vegetation x_A	[0 , 1]
Canopy cover percentage x_{CC}	[0 ,1]
Rainfall intensity (in mm/h) x_I	{10, 25, 30} depending on studied area climate
Hauteur de la vgtation (en m) x_{PH}	[0 , 130]

$$F(x_K, x_R, x_A, x_{CC}, x_I, x_{PH}) = x_k \times [x_R \times x_A \times (1 - x_{CC}) \times (11.9 + 8.7\ log\ x_I)$$
$$+(15.8 + x_{PH}\ ^{0.5}) - 5.87] \times 10^{-3}$$

Raindrop detachment model in RMMF

Deriving new primitive constraints

- Proposition
  - Defining constraints that are derived from expert models
  - Using them during pattern mining to improve both relevancy and scalability thanks to formal properties

- Considered primitive constraint

$$q_{F\geq}(X) \equiv F(X) \geq minf$$

X is an itemset and F is an expert model (a multivariate function)

Minimal erosion constraint

- Minimal erosion in an area $q_{F\geq}(X) \equiv F(X) \geq minf$

Focus on patterns X related to a high soil particle detachment or a high erosion risk

- This constraint can be combined with others, e.g., a minimal frequency
- It enables to prune patterns that are not related to soil erosion
- It highlights patterns validated by the data and the model
- It may support the detection of contradictions w.r.t. expert model output

Value $F(X)$ for an itemset X

Trivial case: itemset X involves all the variables of the model

- Model RMMF $F(x_K, x_R, x_A, x_{CC}, x_I, x_{PH})$

x_K soil detachment index	x_{CC} canopy cover percentage
x_R annual rainfall	x_I rainfall intensity
x_A proportion of rain stopped by vegetation	x_{PH} vegetation height

X={x_K=laterite, x_R=6000, x_A=0.3, x_{CC}=0.1, x_I=25, x_{PH}=1, ... }

Compute F(4, 6000, 0.3, 0.1, 25, 1) and ignore the other items not considered by the model

What is the value associated to an itemset when some of the variables of the model are not involved?

Value $f(X)$ for an itemset X (2)

Bound consistency

Model RMMF $F(x_K, x_R, x_A, x_{CC}, x_I, x_{PH})$

x_K soil detachment index	x_{CC} canopy cover percentage
x_R annual rainfall	x_I rainfall intensity
x_A proportion of rain stopped by vegetation	x_{PH} vegetation height

$Z=\{x_K=$laterite, $x_R=6000$, $x_A=0.3$, $x_{CC}=0.1$, $x_I=25\}$

F(4, 6000, 0.3, 0.1, 25, ?)

Solution: Computing upper and lower bounds for $F(Z)$

$$min(F(4, 6000, 0.3, 0.1, 25, i)) \leq F(Z) \\ \leq max(F(4, 6000, 0.3, 0.1, 25, i)) \forall i \in dom(x_{PH})$$

i.e.,

$$F(Z) = [F(4, 6000, 0.3, 0.1, 25, \mathbf{0}), F(4, 6000, 0.3, 0.1, 25, \mathbf{130})]$$

Properties of $F(X) \geq minf$ (1)

- Studying constraint properties to improve mining scalability
  - Looking for safe pruning rules
  - A classical approach in the pattern mining area: Using monotonicity properties
- Focus on 2 types of properties
  - Properties that can be used to prune supersets
  - A property that can be used to prune "neighborhood" patterns within the search space

Properties of $F(X) \geq minf$ (2)

If $q_{F\geq}(X)$ is false, then all its supersets Y expressing the same variables of the model can be pruned
If $q_{F\geq}(X)$ is false then $q_{F\geq}(Y)$ is false, because $F(X) = F(Y)$

Given $F(X)$ in $[inf_x, sup_x]$. If $F(X) < minf$, i.e., $sup_x < minf$, then all the supersets Y of X can be pruned
If $F(X) < minf$ then $F(Y) < minf$ because $F(X) \geq F(Y)$

Properties of $F(X) \geq minf$ (3)

Assume Z and Y are "direct neighbors"

- They share the same variables of the model
- For these variables, all the values are identical (i.e., items) except for a variable x_j

Assume $Z.x_j, Y.x_j \in [a, b]$, $\frac{\partial F}{\partial x_j} > 0$ on $[a, b]$.
If $F(Z) < minf$ and $Z.x_j > Y.x_j$ then $F(Y) < minf$

N.B.: idem if $\frac{\partial f}{\partial x_j} < 0$ and $Z.x_j < Y.x_j$

Algorithms

Integrating constraint evaluation into itemset mining algorithms
These properties can be directly integrated in existing itemset mining algorithms

- e.g., Apriori [Agrawal and Srikant 94] or Close-by-One [Kuznetsov and Obiedkov 02]
- Integration is more or less easy depending on the enumeration strategy

Empirical validation

Data

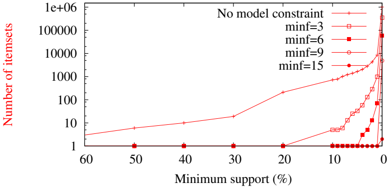
- A SPOT satellite image with 8 millions of pixels
    - 5 discretized radiometric variables (50 items)

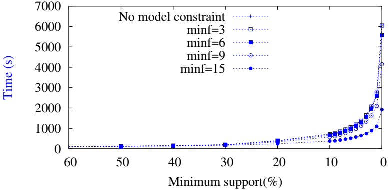
\+ data on nature of the soil, land cover and slope (24 items)

Experimental protocol

- Model from [Atherton et al. 05] that assesses an erosion risk
- Integration in the algorithm Close-By-One
    - Constraint: Looking for frequent and closed itemsets associated to a minimal erosion
- Tests with several frequency thresholds, with and without the model-based constraint (with various model thresholds)

Experimental results (1)





Number of solutions and execution times decrease from 10 000 itemsets in 6000 sec. to 10 itemsets in 2000 sec. (*minsup* $= 10\%$ + no model-based constraint $\rightarrow$ *minf* $= 3$)

Experimental results (2)

Qualitative feedback

- 1% of the studied area is associated to a strong soil erosion risk
- High risk areas are characterized by serpentinite soils covered by volcano-sedimentary substrat and have an important slope
- The results are confirmed by the radiometric values associated with the pattern (low green band and NDVI)

See details in Ph. D thesis by Jérémy Sanhes (September 25, 2014, In French)

Conclusion & perspectives

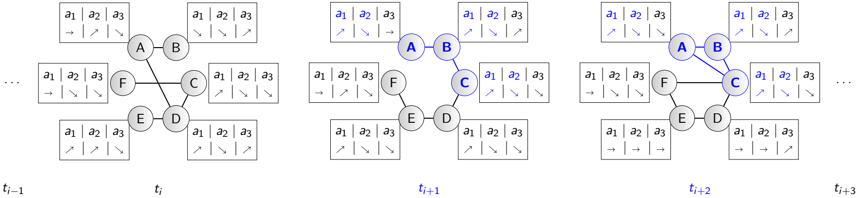
- Problem: Integrating domain knowledge and expert knowledge within constraint-based pattern mining techniques
- Proposition: Deriving new constraints based on available models expressed as multivariate functions
  - Preliminary results in the simple context of itemset mining with an application to soil erosion understanding
  - It improves pattern relevancy and data mining scalability, i.e., supporting knowledge discovery from data
  - It promotes better interactions with experts of the application domain: a needed step towards pattern domain co-design
  - Pattern domain prototyping could be a key methodology for data science and using Constraint Programming for that purpose is obviously promising

$$\{\phi \in \mathcal{L} \mid \mathcal{C}_1(\phi, \mathcal{R}) \wedge \ldots \wedge \mathcal{C}_p(\phi, \mathcal{R}) \wedge \mathcal{C}_{p+1}(\phi, \mathcal{M}, \mathcal{R}) \wedge \ldots\}$$

Conclusion & perspectives

- Some perspectives
  - Defining and studying other model-based primitive constraints (s.t. enforcing unexpectedness)
  - Studying other families of models for other applications (e.g., epidemiology models, sociological models of information diffusion, logistic models)
  - Using this approach with other data mining methods (e.g., dynamic graph mining methods)

Trends in dynamic attributed graphs



$$\left\{\{(A,B,C),(t_{i+1},t_{i+2})\},(a_1^+,a_2^-)\right\}$$

Desmier et al. ECML/PKDD 2013.

Questions ?

Thanks to:

- Loïc Cerf (Belo Honrizonte, UGMG)
- Elise Desmier
- **Frédéric Flouvat** (PPME, Nouméa)
- Marc Plantevit (LIRIS, Lyon)
- Adriana Prado
- Céline Robardet (LIRIS, Lyon)
- **Jérémy Sanhes** (PPME, Nouméa)
- **Nazha Selmaoui-Folcher** (PPME, Nouméa)