Calcul de motifs sous contraintes pour la classification supervisée

Constraint-based pattern mining for supervised classification

Dominique Joël Gay

Soutenance de thèse
pour l'obtention du grade de docteur en informatique

Encadrement : Nazha Selmaoui-Folcher et Jean-François Boulicaut

ERIM EA3791, PPME EA3325, Université de la Nouvelle-Calédonie
LIRIS CNRS UMR5205, INSA-Lyon

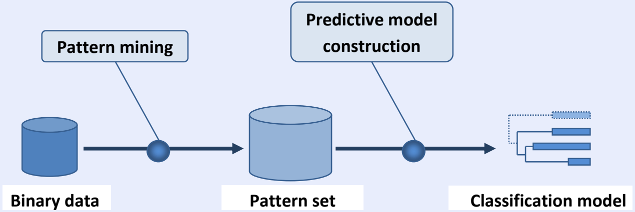
30 novembre 2009

Context

Supervised classification. . .

. . . in labeled 0/1 samples

Recent developments : Pattern-based classification



Challenging problems

- Classification in noisy data
- Classification in multi-class imbalanced data

Contributions to open problems

Classification when attributes are noisy

an application-independent pattern-based noise-tolerant feature construction method

Multi-class imbalanced classification

- a new framework dedicated to multi-class imbalanced data
- a parameter-free pattern-based method

Plan

1. Preliminaries
2. NTFC : Noise-Tolerant Feature Construction
3. Multi-class imbalanced classification : `fitcare`
4. Application to soil erosion characterization
5. Conclusion & Perspectives

Pattern-based classification : an example

Shall we organize "Les jeudis du centre ville" if it's rainy, with cooling temperature and without wind ?

"Les jeudis du centre-ville"

r		Attributes										Classes
		outlook			temperature			humidity		windy		jeudi
		sunny	overcast	rainy	hot	mild	cool	high	normal	true	false	yes/no
Objects (Training)	t_1	1	0	0	1	0	0	1	0	0	1	no
	t_2	1	0	0	1	0	0	1	0	1	0	no
	t_{14}	0	0	1	0	1	0	1	0	1	0	no
	t_8	1	0	0	0	1	0	1	0	0	1	no
	t_3	0	1	0	1	0	0	1	0	0	1	yes
	t_5	0	0	1	0	0	1	0	1	0	1	yes
	t_7	0	1	0	0	0	1	0	1	1	0	yes
	t_9	1	0	0	0	0	1	0	1	0	1	yes
	t_{10}	0	0	1	0	1	0	0	1	0	1	yes
	t_{11}	1	0	0	0	1	0	0	1	1	0	yes
	t_{12}	0	1	0	0	1	0	1	0	1	0	yes
	t_{13}	0	1	0	1	0	0	0	1	0	1	yes
Test	t_4	0	0	1	0	1	0	1	0	0	1	?
	t_6	0	0	1	0	0	1	0	1	1	0	?

Pattern mining and classification

Task

Mining a set of relevant class-characterizing patterns to predict class labels

Various types of pattern

- Association rules (γ-frequency , confidence)(Agrawal et al. SIGMOD'93)
  π : outlook_sunny and humidity_normal $\rightarrow$ yes
  (freq : 2 ; conf : 1)
- Emerging itemsets (γ-frequent ρ-EPs)(Dong et al. KDD'99)
  humidity_high $\rightarrow$ no
  (freq : 6 ; GR : 4)
- inductive rules, . . .

How to predict class labels for a new incoming object t ?

Combining patterns supported by t to compute a score.

Plan

Noise handling : what has been done ?

Effects of noise

- Class-noise / Attribute-noise
- Low classification performance / low accuracy results

Noise handling

- Class-noise / Attribute-noise
- Noise detection / filtering / deletion / correction

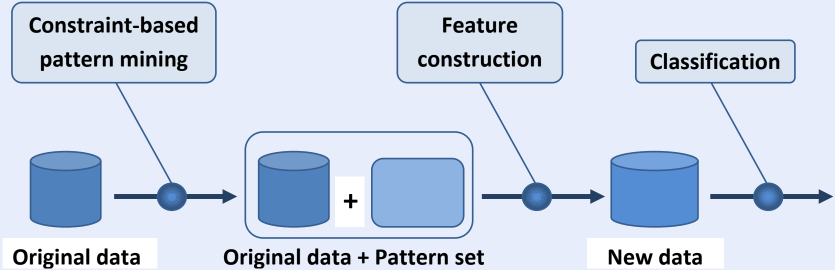
↪ Undesirable information loss

Our proposal

- Robust (noise-tolerant) feature construction based on frequent patterns
- without filtering, deleting or correcting any instance

Our proposal

Noise-Tolerant Feature Construction processus (NTFC)



A relevant pattern is . . .

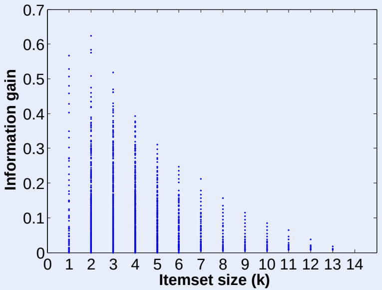
- frequent itemset
- class-characterizing
- noise-tolerant

Why itemsets ? Why frequent ones ?

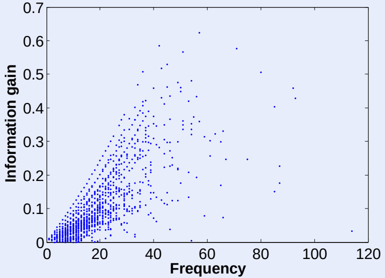
Intuition

"A frequent itemset could be interesting"

Interestingness of k-itemsets



Interestingness of frequent itemsets



Frequent itemsets are preferable to single items

Frequent itemsets : which ones ?

Pattern-based classification : key points

Let Y be an itemset characterizing class c_i.

- Discrimination (w.r.t. an interestingness measure)

Let S be a set of itemsets characterizing class c_i.

- Coverage of training data ($\sim$ for a relevant data set representation)
- Minimality : $\nexists X$ characterizing c_i s.t. $X \subseteq Y$
- Redundancy : $Z \in S$ characterizing c_i s.t. $Y \subseteq Z$ is redundant
- S is a concise set

Redundancy has been studied by means of the so-called condensed representations of frequent itemsets

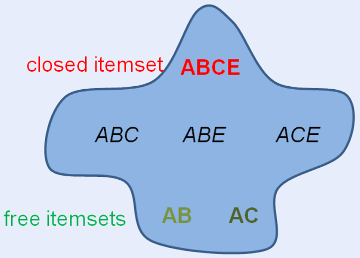
Closure Equivalence Classes (CECs)

Bastide et al. SIGKDD Expl.'00 / Boulicaut et al. PKDD'00

Grouping itemsets having the same support/closure (CECs)

r	A	B	C	D	E	F
t_1	1	1	1	1	1	0
t_2	1	1	1	0	1	0
t_3	0	1	1	0	1	0
t_4	1	0	0	1	1	0
t_5	0	1	1	0	0	1
t_6	0	1	0	1	0	1

$\gamma = 2$
$freq(AB) = freq(ABCE) = 2$ (equivalent support)
$cl(AB) = cl(AC) = cl(ABC)$
$= cl(ABE) = cl(ACE)$
$= cl(ABCE) = ABCE$



Important

From a CEC, we may derive a strong association rule between a free itemset and each element of its closure
e.g., $AB \Rightarrow C$ and $AB \Rightarrow E$ are strong rules.

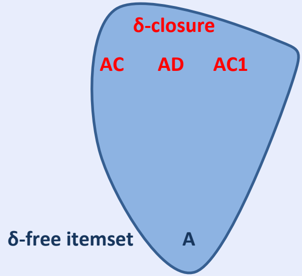
δ-Closure Equivalence Classes (δ-CECs) Boulicaut et al. DMKD'03

From equivalence classes (δ-CECs) to relevant itemsets

r	A	B	C	D	c_1	c_2
t_1	1	1	1	1	1	0
t_2	1	1	1	0	1	0
t_3	0	1	1	0	1	0
t_4	1	0	0	1	1	0
t_5	0	1	1	1	0	1
t_6	0	0	1	1	0	1
t_7	1	0	1	1	0	1

$\gamma = 3; \delta = 1$
$freq(A) = 4$
$freq(AC) = 3$
$freq(AD) = 3$

$freq(Ac_1) = 3$

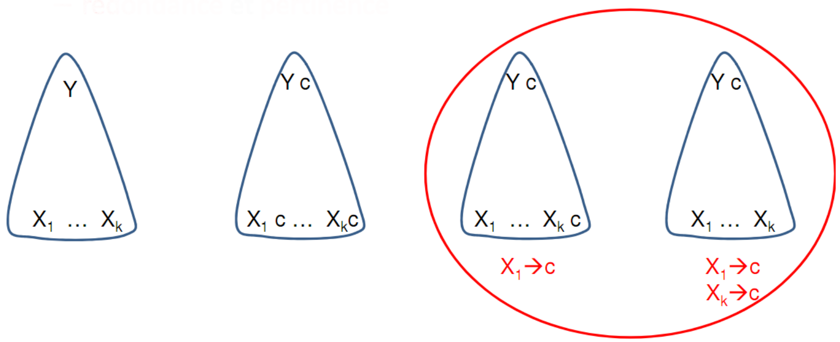


Important

From a δ-CEC, we may derive a δ-strong rule between a δ-free itemset and each element of its δ-closure.
e.g., $A \Rightarrow D$ and $A \Rightarrow c_1$ are δ-strong rules.

Which δ-CECs?

Various types of δ-CECs...



We may derive one or more relevant (γ-frequent δ-free) itemsets from a δ-CEC.

Fair properties of δ-CECs Crémilleux et al. ES'02

Let $\pi : X \rightarrow c$ be a δ-strong rule (X is a γ-frequent δ-free itemset).

High confidence

For small δ values, π is highly confident : $conf(\pi, r) \geq 1 - \delta/\gamma$

Avoiding classification conflicts

If $\delta < \gamma/2$, we may avoid following conflicts :

- equal body conflict ($\pi' : X \rightarrow c'$ does not exist)
- included body conflict ($\pi' : X \cup Y \rightarrow c'$ does not exist)

Discriminative power of Emerging patterns Gay et al. PAKDD'08

If $\delta < \frac{\gamma \cdot |r \backslash r_{c_i}|}{\rho \cdot |r|}$ (r_{c_i} is the major class), X is a ρ-emerging pattern.

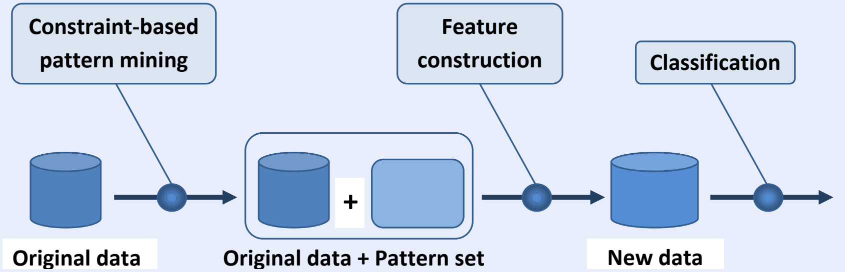
$\hookrightarrow$ Set $S_{\gamma,\delta}$ of relevant non-conflicting δ-strong characterization rules

Feature construction process

Principle

- From each itemset (rule body), a new descriptor is built.
- For $\pi : X \rightarrow c$, we have $\mathrm{NewAttribute}(t, X) = \frac{|X \cap \mathrm{Items}(t,X)|}{|X|}$
- Thus, $\mathrm{NewAttribute}(t) \in \{0, \frac{1}{|X|}, \ldots, \frac{|X|-1}{|X|}, 1\}$

Noise-Tolerant Feature Construction processus (NTFC)

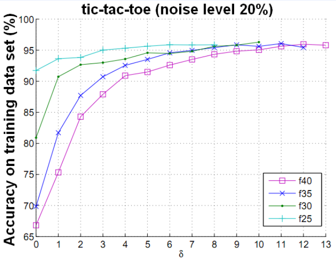


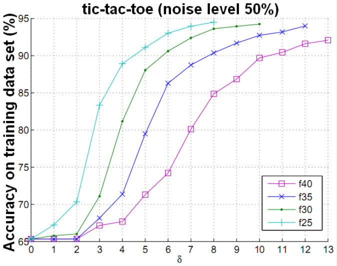
Experiments : Impact/tuning of parameters γ and δ

Minimum frequency : γ

- Extreme values $\Rightarrow$ low interest
- Tuning ? Still an open question. . .

A strategy for δ setting (given γ)





- Increasing δ starting from 0 until stability point : δ_{opt}

Experimental protocol

Protocol

1. UCI data sets
2. Uniform attribute-noise injection [0-50%] only in training data sets
3. Classification techniques : C4.5, NB and SVM

- 11 data sets : 55 noisy versions
- Two types of accuracy results : using δ_{opt} and using the best γ, δ combination

Experimental results

Data sets enhanced by NTFC versus Original data sets

NTFC-C4.5 vs C4.5

- Best : 50 / 5
- δ_{opt} : 35 / 20

NTFC-NB vs NB

- Best : 41 / 14
- δ_{opt} : 28 / 27

NTFC-SVM vs SVM

- Best : 42 / 13
- δ_{opt} : 40 / 15

NTFC vs HARMONY

NB, C4.5, SVM $<$ HARMONY $\leq$ NTFC-NB, NTFC-C4.5, NTFC-SVM

Summary

Summary

+ A solution to deal with attribute-noisy data : enhancing data set description with robust features
- For imbalanced data sets a low γ threshold is needed
  It enforces low δ values. . .

Plan

Multi-class imbalanced problem with L.Cerf

Effects of class disproportion

- Low per class accuracy results for minor class(es)
- Bias towards the majority class

Handling imbalanced problems

- Re-balance class distribution by over/under-sampling

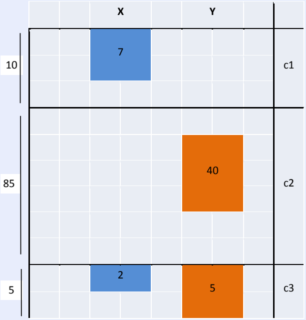
$\hookrightarrow$ Under-sampling : Undesirable information loss
$\hookrightarrow$ Over-sampling : Over-fitting, additional computational task

Pattern-based classification

Handling multi-class imbalanced problem with pattern-based techniques ?

Limits of existing frameworks

Examples of One-Versus-All (OVA) frameworks



- Frequency-confidence framework
  $Conf(Y \rightarrow c_2, r) = 40/45$
  $GR(Y, r_{c_3}) = \frac{5/5}{40/95} > 2$
  Y characterizing class c_2 or c_3 ?
- EPs-based framework
  $GR(X, r_{c_1}) = (7/10)/(2/90) > 31$
  $GR(X, r_{c_3}) = (2/5)/(7/95) > 5$
  X characterizing class c_1 or c_3 ?
- Positive correlation framework
  $FInt(X, c_1, r) = (100 \times 7)/(9 \times 10) > 1$
  $FInt(X, c_3, r) = (100 \times 2)/(9 \times 5) > 1$
  X characterizing class c_1 or c_3 ?

Other Limits of existing frameworks

Unsuitable global frequency threshold

Frameworks using a global frequency threshold are biased towards the majority class

Causes of limitations

- Class distribution is not taken into account
- Repartition of errors made by pattern into classes is not taken into account

Idea

`One-Versus-Each` framework :

- having a frequency threshold per class
- for each class, having an error threshold per each other class

OVE framework : OVE-characterizing rules (OVE-CRs)

OVE-characterizing rules

An OVE-characterizing rule for a class c_i is :

- frequent in r_{c_i} (relatively)
- infrequent (relatively) in every other class taken independently
- as general as possible

OVE-characterizing rule (formally)

An OVE-characterizing rule for r_{c_i} is :

- $freq_r(X, r_{c_i}) \geq \gamma_{i,i}$
- $\forall j \neq i, freq_r(X, r_{c_j}) < \gamma_{i,j}$
- $\forall Y \subset X, \exists j \neq i \mid freq_r(Y, r_{c_j}) \geq \gamma_{i,j}$

OVE framework : matrix of parameters and constraints

p^2 parameters for a p-class problem

$$\Gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,p} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p,1} & \gamma_{p,2} & \cdots & \gamma_{p,p} \end{pmatrix}$$

Consistency constraints on Γ lines / columns

$$\mathbb{C}_{line} \equiv \forall i \in \{1, \ldots, n\}, \forall j \neq i, \gamma_{i,j} < \gamma_{i,i}$$

$$\mathbb{C}_{column} \equiv \forall i \in \{1, \ldots, n\}, \forall j \neq i, \gamma_{j,i} < \gamma_{i,i}$$

Conflictless rule set

$\mathbb{C}_{line} = true \wedge \mathbb{C}_{column} = true \Rightarrow$ rule set S_Γ of OVE-CRs is conflictless.

OVE framework : fitcare

Our proposal

fitcare : an OVE parameter-free associative classification method

- Extraction of a set S_Γ of OVE-CRs w.r.t. Γ
- Classification based on S_Γ
- Automatic parameter tuning (locally optimal consistent parameters Γ_{opt})

Efficiently mining OVE-CRs and classification

Extraction

- Breadth-first search strategy
- anti-monotonicity properties of minimum frequency and minimal body constraints
- Per-class mining : $S = \cup_{i \in \{1,\dots,p\}} S_{c_i}$

Classification using per class frequencies

Given an object t, its likeliness score in c_i is :

$$l(t, c_i) = \sum_{\{c \in \mathcal{C}\}} \left(\sum_{\{\pi : X \rightarrow c \in S | X \subseteq Items(t,r)\}} freq_r(X, r_{c_i}) \right)$$

The class which maximizes l, indicates the class label for t.

`fitcare` : an optimization-based method

Principle : Hill-Climbing

Maximizing the quality of the classifier (rule set) by adjusting the most promising parameter of Γ with commit/roolback strategy.

Initialization

- reaches the first stable state for Γ w.r.t. $\mathbb{C}_{line}$ and $\mathbb{C}_{column}$
- indicates the maximal positive cover rate obtained.

Positive cover rate

The maximal positive cover rate obtained at initialization must be maintained during the optimization phase.

fitcare : optimization

Confusion measure : Objective function to optimize

- Measuring the confusion made with class c_j when classifying objets of $\mathcal{T}_{c_i}$.

$$g(c_i, c_j) = \frac{\sum\limits_{t \in \mathcal{T}_{c_i}} l(t, c_i)}{\sum\limits_{t \in \mathcal{T}_{c_i}} l(t, c_j)}$$

  When g increases, confusion weakens.
- The greatest term of the denominator of g indicates the parameter to lower.

Experiments : accuracy results

Accuracy comparisons (Win-Tie-Loss) with HARMONY and CPAR

Global accuracy

fitcare vs HARMONY : 6-2-11

fitcare vs CPAR : 14-1-4

Per class accuracies for minor classes

fitcare vs HARMONY : 13-3-3

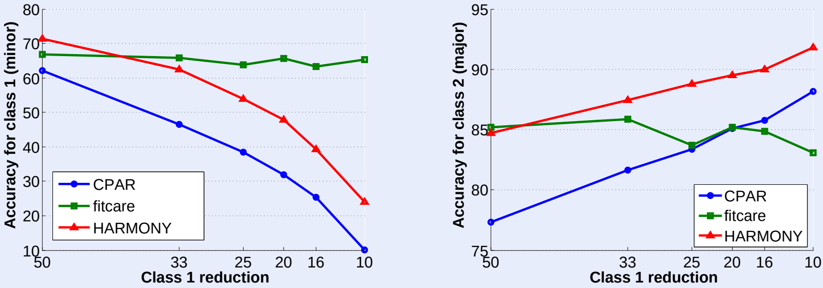
fitcare vs CPAR : 12-4-3

fitcare : performances

OVE fitcare >> OVA HARMONY, CPAR

Experiments : bias towards the majority class

Evolution of accuracies w.r.t. class reduction



Summary

OVE `fitcare` avoids the bias towards the majority class

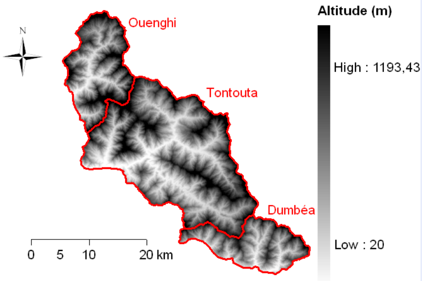
Plan

Soil erosion characterization with I.Rouet

Tasks

- Combinations of attributes that are suitable for erosion phenomenon ?
- Semi-automatic mapping of erosion zones in a region.
- What about erosion hazard ?

3 catchment basins

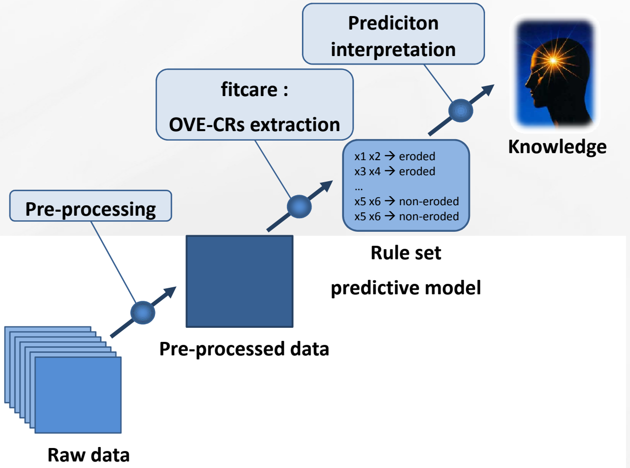


Various informations (per pixel)

- Rain fall
- Lithology
- Altitude
- Land cover
- Slope
- *Erosion* : (eroded soil / non-eroded soil)

Knowledge discovery process . . .

. . . in erosion data set



Results : analysis of OVE-CRs set

Soil erosion characterization

A set of OVE-CRs confirmed by domain experts. Now, observed phenomenons may be quantified (with frequency values) and qualified (with Growth rate values).

Combinations favourable to (non) appearance of erosion

exploitation and excavation mining zones → eroded soil (0.0280 ; 0,0008 ; 31.6)

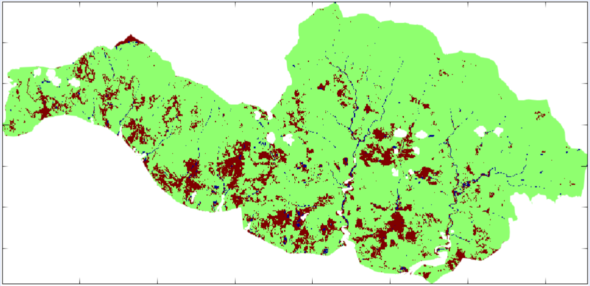
- Frequency in $r_{eroded} = 0.0280$
- Frequency in $r_{non-eroded} = 0.0008$
- Growth rate $= 31.6$

dense forest → non-eroded soil (0.0114 ; 0.0902 ; 7.9)

...

Results : prediction

Semi-automatic mapping



Confusion matrix

Dumbea Predictions	Real classes non-eroded	Real classes eroded
non-eroded	112827	743
eroded	14437	**926** $\simeq 55\%$

Ouenghi Predictions	Real classes non-eroded	Real classes eroded
non-eroded	137016	835
eroded	31173	**3194** $\simeq 79\%$

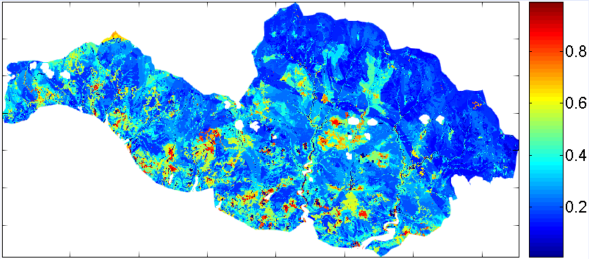
Global accuracy : `fitcare` $\simeq$ `C4.5`, `NB`, `HARMONY`

Accuracy for minor class : `fitcare` $\gg$ `C4.5`, `NB`, `HARMONY`

Results : erosion hazard

Erosion hazard estimation

Estimation of the probability of erosion occurrence using per class frequencies normalization (with `fitcare`)



Plan

Conclusion & Perspectives

Summary

Contributions to open questions : pattern-based classification in difficult contexts

Noisy data

- A generic robust feature construction method
- New NTFC features $\gg_{better}$ original features

Noisy data : perspectives

- What about class-noise handling ?

Conclusion & Perspectives

`fitcare` : bilan

- A new framework and method dedicated to multi-class imbalanced problem
- High per class accuracies for minor classes
- Solving the problem of bias towards the majority class

`fitcare` : perspectives

- Exploring the field of optimization algorithms, local optimum search ...
- Cost-sensitive `fitcare` ?

Conclusion & Perspectives

Application to erosion data set

- Quantification and qualification of erosion phenomenon
- Semi-Automatic mapping of erosion zone in a region
- Erosion hazard assessment

Applications : perspectives

- Generic methodological contributions $\Rightarrow$ Various applications
- From pixel data to spatial data and spatial pattern mining

That's all folks !

Question time

Thank you for your attention.