LENA, La Pitié Salpétrière

22 mars 2005

Fouille de Données MEG et Optimisation multi-critères

Michèle Sebag

IA – TAO, CNRS – INRIA

Université Paris-Sud Orsay, `http://tao.lri.fr`

Travail joint : Nicolas Tarrisson, Olivier Teytaud, Sylvain Baillet, Julien Lefevre.

Contexte

Neuro-imagerie

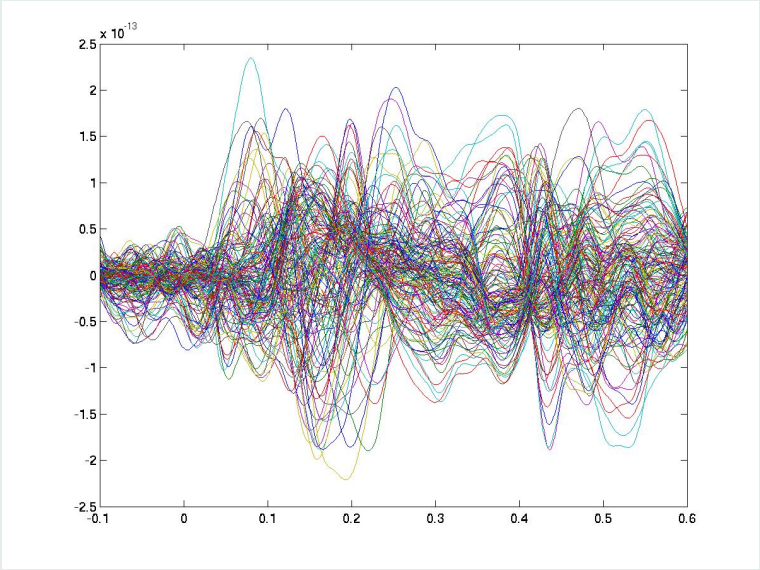
- Des sujets, une expérience, des mesures

- Une nouvelle technologie : électro-magnéto-encéphalographie.
  pas de temps: .001 seconde

Plan

- Le problème posé
- Fouille de données
- Optimisation multi-critères
- .. multi-modale
- 4dMiner
- Discussion

Les données



Structure spatio-temporelle

- Capteurs $i = 1..N$
- $i \quad \rightarrow \quad \begin{cases} M_i = (x_i, y_i, z_i) \ \in \mathbb{R}^3 \\ \{C_i[t], t = 1..T\} \ \in \mathbb{R}^T \end{cases}$

Le but

Trouver des motifs spatio-temporels

- Une aire spatiale A : Boule $\mathcal{B}$(centre i_0, rayon r)
- Un intervalle temporel $I \subset \{1..T\}$ caractérisant

$$\mathcal{V}(A, I) = \{C_k[t],\ k\ \in\ A,\ t\ \in I\}$$

TELS QUE

- A large
- I large
- Variance $(\mathcal{V}(A, T))$ faible

Etat de l'art

Procédure courante

- manuelle $\rightarrow$ i) ennuyeux; ii) subjectif
- peu de volontaires.

Analyse en composantes indépendantes

- Identifier les sources i) pb inverse, ii) hyp. sur la nature des sources
- Identifier les périodes de stabilité des sources.

Champs conditionnels de Markov

- Idem, dans un espace différent.

Fouille de Données

Mot d'ordre

- à partir de données *et de connaissances*
- fournir à l'expert des régularités utiles, nouvelles, valides

Contexte

L'idéal	le siècle des connaissances
La réalité	des expertises spécialisées
Le besoin	la gestion humaine des connaissances ne passe pas à l'échelle
L'opportunité	les données sont accessibles

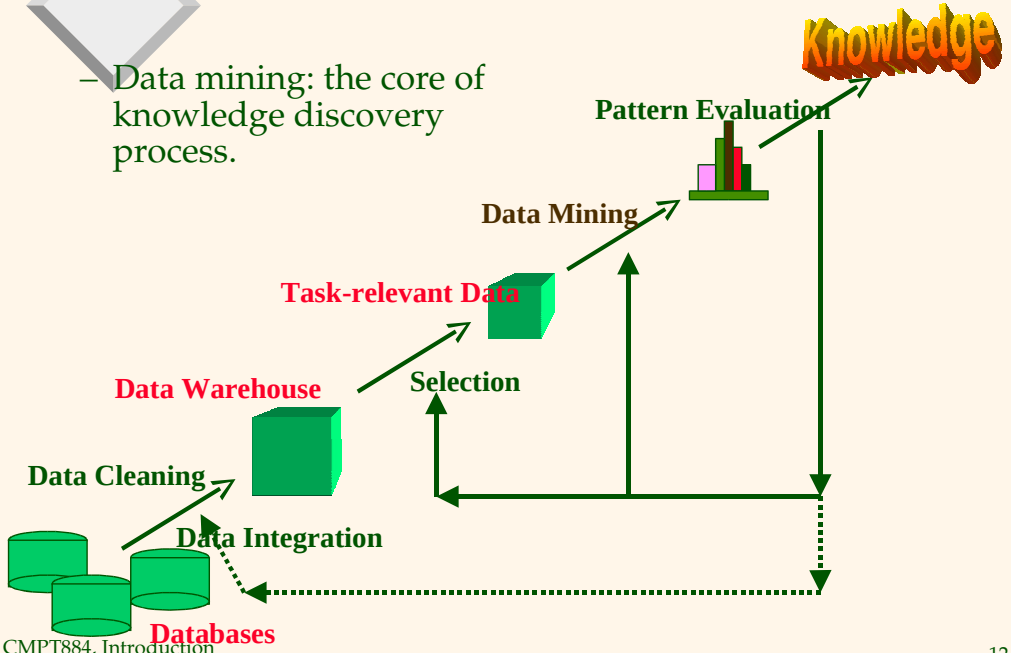
Domaines d'application

Domaine	**But : Modélisation**
Phénomènes physiques	modélisation et contrôle de process
Applications industrielles, sciences expérimentales, calcul numérique	
Phénomènes sociaux	+ confidentialité
Hôpitaux, Assurances, Banques, ...	
Phénomènes individuels	+ dynamique rapide
Consumer Relationship Management	
User Modelling	

NoE PASCAL : http://www.pascal-netwwork.org

Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



Fouille de Données, Résumé

Une définition... Fayyad et al. 1996

Automatic extraction of
novel, useful and valid knowledge
from large sets of data.

...faisant une large part à l'implicite...

des connaissances { nouvelles (% au sens commun) ; utiles (pour qui) ; valides (un pb multi-critères)

Un cahier des charges :

- Idéalement, un système adaptatif
- Doit passer à l'échelle

Fouille de MEG : une approche algorithmique

Propriétés voulues

- Passage à l'échelle
- Flexible $\Rightarrow$ Paramétrable $\Rightarrow$ Doit être calibré

$\Rightarrow$ Contrôle des ressources possible - Algorithme anytime
`http://anytime.cs.umass.edu/~shlomo/`

Discussion

- critères monotones (en r, en I)
- critères antagonistes ($I \nearrow$, $r_A \searrow$)
- exhaustivité ? Non : le résultat doit être vu par un humain.

Optimisation multi-critères

Optimisation classique

Trouver $ArgMax\{\mathcal{F}(x), \mathcal{F} : \Omega \to \mathbb{R}\}$

Optimisation multi-critères

Trouver $ArgMax\{\mathcal{F}_i,\ i = 1, 2..., \mathcal{F}_i : \Omega \to \mathbb{R}\}$

Evidemment, $\mathcal{F}_i$ antagonistes.

De qualité maximale, de prix minimal...

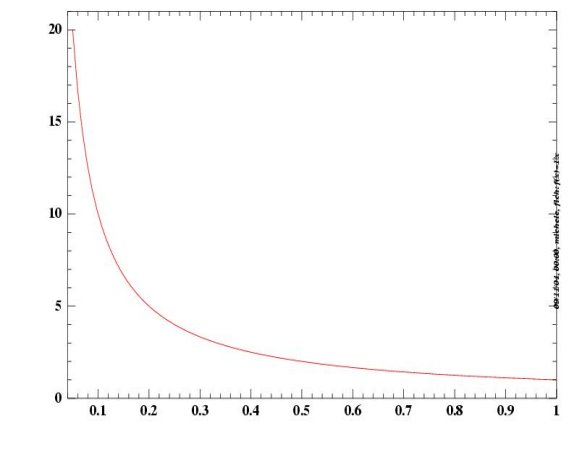
Front de Pareto

Domination de Pareto

- $x < y$ ssi $\mathcal{F}_i(x) \leq \mathcal{F}_i(y)$ et inégalité stricte pour au moins un i.

Front de Pareto

- Ensemble des solutions non dominées.



Stable Spatio-Temporal Patterns

Espace de recherche

$$X = \begin{cases} I & \text{intervalle temporel} \\ i & \text{centre de la boule spatiale} \\ r & \text{rayon de la boule spatiale} \\ d_w = (a, b, c) & \text{distance pondérée} \end{cases}$$

avec *régions ellipsoidales*

$$d_w(j,k) = a.(x_j - x_k)^2 + b.(y_j - y_k)^2 + c(z_j - z_k)^2$$

Objectifs d'optimisation

$X = (I = [deb, fin],\ i \text{ centre},\ d_w = (a, b, c),\ r)$

Définition

I-alignement de deux capteurs j et k sur l'intervalle temporel I:

$$\sigma_I(j,k) \ = \ < j,k >_I \times \left(1 - \frac{|\bar{C}_j^I - \bar{C}_k^I|}{|\bar{C}_j^I|}\right)$$

avec

$$< j,k >_I \ = \ \frac{\sum_{t=t_1}^{t_2} C_j(t).C_k(t)}{\sqrt{\sum_{t=t_1}^{t_2} C_j(t)^2 \ \times \ \sum_{t=t_1}^{t_2} C_k(t)^2}}$$

$$\bar{C}_j^I = \ \text{Moyenne}\left\{C_j[t],\ t \in I\right\}$$

Objectifs d'optimisation, II

$X = (I = [deb, fin], i \text{ centre}, d_w = (a, b, c), r)$

Objectifs

- Longueur temporelle $\ell(X) = fin - deb$
- Le voisinage spatial $\mathcal{V}(X) = \{j \ / \ d_w(i, j) < r\}$
- La taille $|X| = \ell(X) \ . \ |\mathcal{V}(X)|$
- La cohérence spatio-temporelle

$$\sigma(X) = \frac{1}{a(X)} \sum_{j \in \mathcal{B}(i,w,r)} \sigma_I(i, j)$$

Algorithmes d'Evolution

$\mathcal{F} : \Omega \mapsto \mathbb{R}$; Trouver les optima de $\mathcal{F}$

Métaphore : L'évolution darwinienne des populations biologiques.

Les individus les plus adaptés survivent et se reproduisent

Vocabulaire :

Individu	Elément X de Ω
Performance	Valeur de $\mathcal{F}(X)$
Population	Ensemble de P éléments de Ω
Génération	Passage de la population Π_i à Π_{i+1}

Processus :

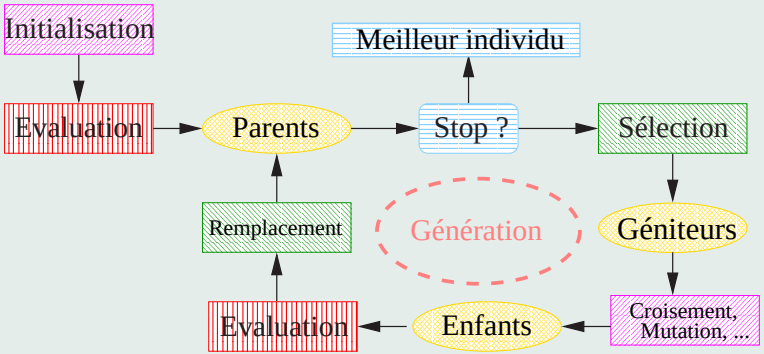
1) Sous la pression du milieu,

2) Les individus se croisent, mutent et se reproduisent.

3) Au bout d'un nombre certain de générations, les individus les plus performants apparaissent dans la population.

$\equiv$ les **optima** de $\mathcal{F}$...

Algorithmes d'évolution : Le Squelette



Opérateurs stochastiques: Dépendent de la représentation

"Darwinisme" (stochastique ou déterministe)

Coût calcul

Critère d'arrêt, statistiques, ...

EC Multi-critères

EC classique — Trouver $ArgMax(\mathcal{F})$

- Initialisation
- Variations (croisement, mutation)
- Sélection

Problème multi-critères

Trouver $X = ArgMax\{a(x), \ell(X), \sigma(X)\}$

Modifications essentielles

But : couvrir le front de Pareto — Archive

Sélection : d'après $\mathcal{F}'(X)$, où $\mathcal{F}'$ mesure :

- Le rang de Pareto de X dans la population courante
- Le pourcentage de l'archive dominé par X
- ...

4d Miner

Components

- Several objectives a, ℓ, σ
- Evolutionary Computation on $\Omega =$

$$\{X = (i, w, I, r),\ i \in [1, N], w \in \mathbb{R}^3,\ I \subset [1, T],\ r \in \mathbb{R}\}$$

  - Initialization — sampling mechanism
  - Variation operators
  - Selection

Sampling mechanism $X = (i, w, I, r)$

- i : uniformly drawn in $[1, N]$;
- $w = (1, 1, 1)$ *initial = Euclidean*
- $I =$
  - deb : uniformly drawn in $[1, T]$
  - $\ell(I)$ drawn $\sim \mathcal{N}(min_\ell, min_\ell/10)$ min_ℓ *user supplied*
  - reject if $deb + \ell(I) > T$
- r : such that the ball contains all neighbors with bounded I-alignment:

$$r = min_k\{d_w(i,k) \; s.t. \; \sigma^I_{i,k} > min_\sigma)\}$$

- reject if $a(X) = |\mathcal{B}(i,r)| < min_a$ min_a *user supplied*

Complexity: $\mathcal{O}(N \log N \times min_\ell)$

Variation operators

Mutation $X = (i, w, I, r)$

- Self adaptive mutation of w, r
- Specific mutation operators for i and I.
- Random (initialisation operator)

Crossover $X = (i, w, I, r) \times Y = (i', w', I', r')$

- Restricted mating:

if the spatio-temporal areas are “close enough” *user-supplied*

Selection : Pareto Archive

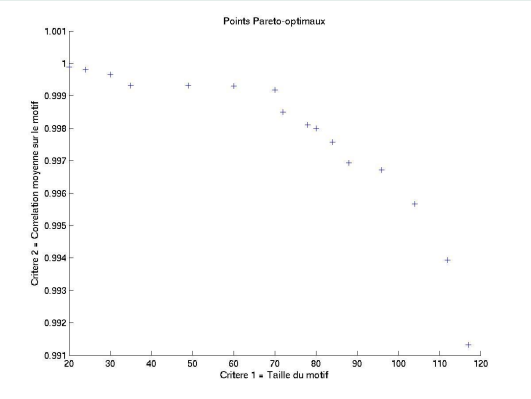
Steady state

- In each step, select an individual (tournament wrt Archive)
- Apply crossover or mutation
- Evaluate
- If non dominated in the population, store :
- Replace an individual (anti-selection)

First results: Failure

Diversity of st-patterns

- seems OK on the Pareto front:

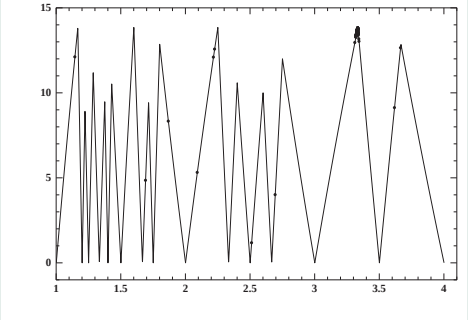


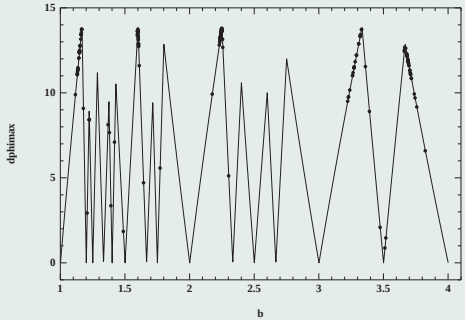
- but all patterns represent the same spatio-temporal region, with variations...

Multi-objective multi-modal optimization

Multi-modal optimization

Find all global *and possibly local* optima





W/o sharing

With sharing

Multi-objective multi-modal optimization

Sharing-Pareto dominance

$X = (i, I, w, r)$ sp-dominates $Y = (i', I', w', r')$ iff

- X Pareto dominates Y wrt a, ℓ, σ
- X and Y overlap $\mathcal{B}(i, w, r) \cap \mathcal{B}(i', w', r') \neq \emptyset$
  $I \cap I' \neq \emptyset$

Experimental validation

Goals of experiment

- Useful ?
- Scalable ?
- Performance / Recall ?

Datasets

- ACI Neurodyne
- Artificial datasets

Artificial datasets

Curves

$N = 500, ..4\ 000$ nb sensors

$T = 1\ 000, ..8\ 000$ nb time steps

For $i = 1..N$

  For $t = 1..T$

    $C_i(t) = C_i(t-1) + \epsilon * \pm 1$

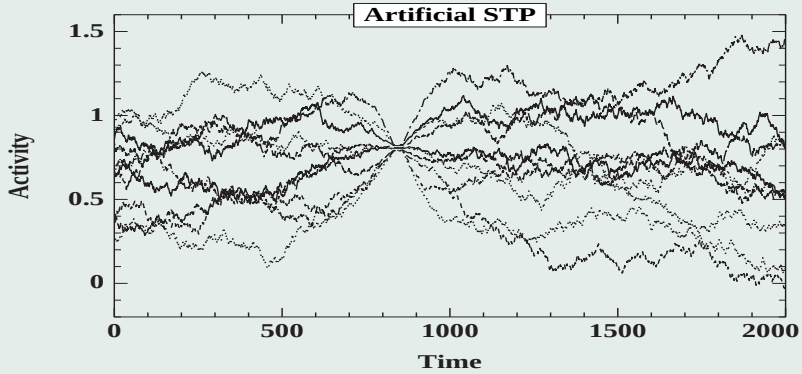
10 Target patterns

P = (i in $1..N$; $I \subset [1, T]$; $w \in \mathbb{R}^3$; $r \in \mathbb{R}$).

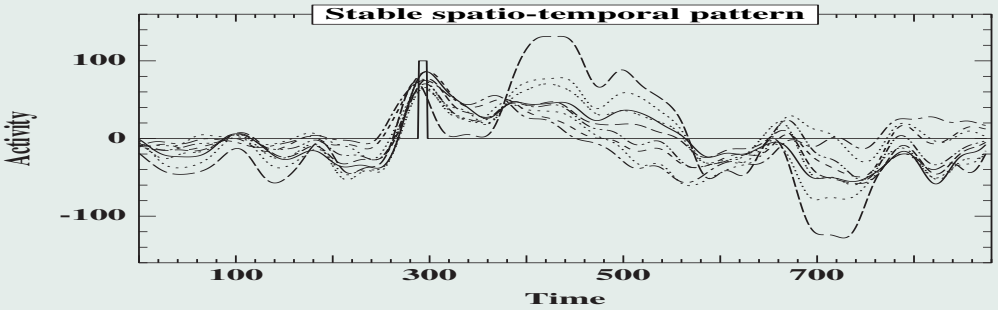
C_P = average of $\{C_j(t), t \in I, d_w(i, j) < r\}$

Action

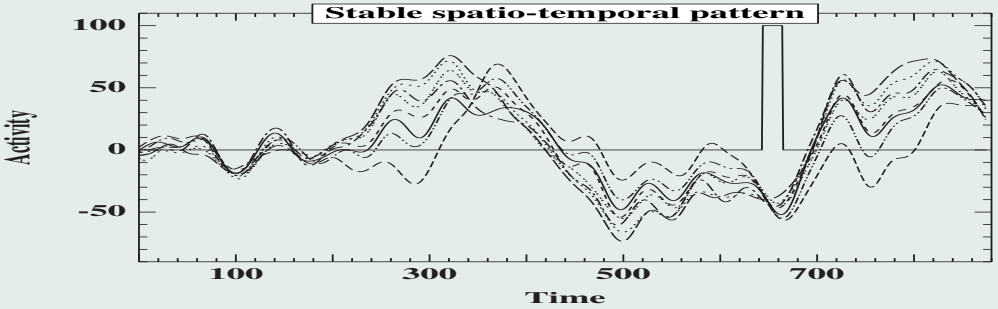
$$C_j(t) = (1 - \alpha)C_j(t) + \alpha C_P \times exp(-d(t, I) - d(i, j))$$


Artificial STP
1.5
1
0.5
0
Activity
0
500
1000
1500
2000
Time

Experimentations MEG



Experimentations MEG

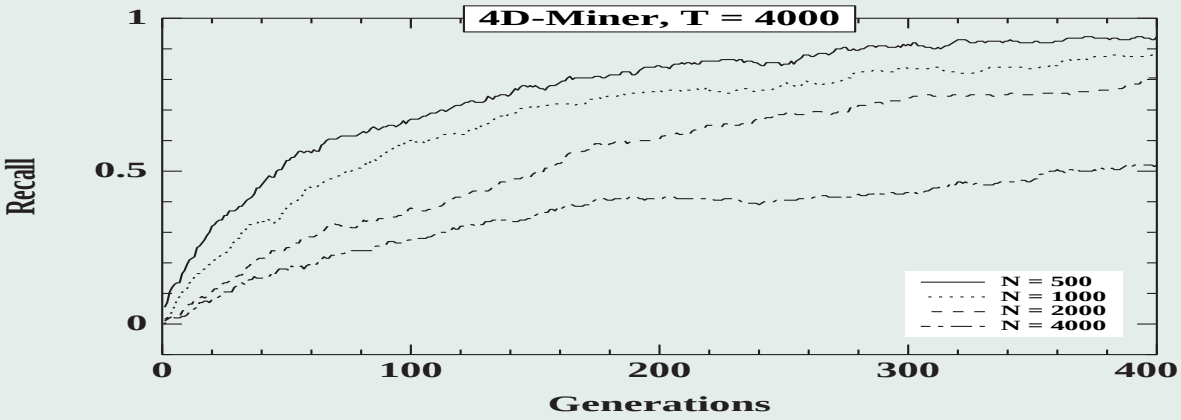


Performances

Recall : percentage of target patterns with representants in the archive.

N	T			
	1,000	2,000	4,000	8,000
500	98 ± 5	93 ± 9	92 ± 7	79 ± 16
1000	96 ± 6	96 ± 6	82 ± 14	67 ± 12
2000	96 ± 5	87 ± 12	72 ± 14	49 ± 15
4000	89 ± 10	81 ± 13	56 ± 14	32 ± 16

Online Performance



Discussion

Convergence

Elagage

Stage DEA Grammaire

BCI