Laboratoire LIPN
Sciences des données

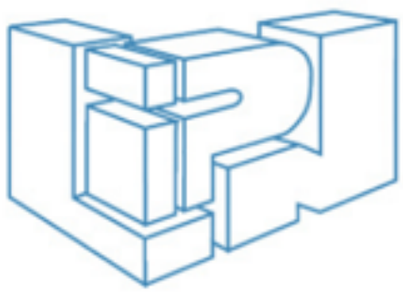
Laure Petrucci

Directrice du Laboratoire LIPN - Univ. Paris 13

Yann Chevaleyre

resp. de l'équipe Apprentissage Artificiel

Laboratoire LIPN - Univ. Paris 13



Structure du LIPN (CNRS UMR 7030)

147 membres dans 5 équipes

- **81 chercheurs**
  1 DR 6 CR
  20+3 PR 50+1 MCF
- **9.5 ITA/BIATSS**
  7+1 CNRS 1.5 Université
- **14 post-doc**
- **42 doctorants**
- plus de 32 stagiaires 15 professeurs visiteurs (en 2015)

Date

Structure du LIPN (CNRS UMR 7030)

147 membres dans 5 équipes

- **81 chercheurs**
  1 DR 6 CR
  20+3 PR 50+1 MCF
- **9.5 ITA/BIATSS**
  7+1 CNRS 1.5 Université
- **14 post-doc**
- **42 doctorants**
- plus de 32 stagiaires 15 professeurs visiteurs (en 2015)

Projets depuis 2012

- 14 ANR, 6 FUI/FEDER/PIA, 2 SPC
- 7 projets collaboratifs internationaux
- 6 projets industriels, 6 CIFRE

Équipes de recherche (1/2)

A^3 : Apprentissage Artificiel et Applications

- Apprentissage et action
- Transformation de l'espace de description pour l'apprentissage par transfert
- Apprentissage non-supervisé massivement distribué

AOC : Algorithmes et Optimisation Combinatoire

- Optimisation dans les graphes
- Programmation mathématique
- Algorithmes, logiciels et architecture distribués

CALIN : Combinatoire, ALgorithmique et INteractions

- Analyse d'algorithmes et structures combinatoires
- Physique combinatoire

Équipes de recherche (2/2)

LCR : Logique, Calcul et Raisonnement

- Logique linéaire et théorie du calcul
- Spécification et vérification modulaires et distribuées

RCLN : Représentation des Connaissances et Langage Naturel

- Analyse de corpus et annotation sémantique
- Découverte et structure des connaissances du web sémantique
- Découverte d'informations sémantiques
- Intégration syntaxe/discours

Implication dans des structures

Fédération de Recherche MathSTIC, FR3734 (Christophe Fouqueré)

- 3 laboratoires :
  - LAGA, UMR 7539 (Philippe Souplet)
  - LIPN, UMR 7030 (Laure Petrucci)
  - L2TI, EA 3043 (Azeddine Beghdadi)
- 3 axes principaux :
  1. Optimisation et apprentissage appliqués aux contenus numériques
  2. Calcul haute-performance, systèmes distribués
  3. Physique mathématique, physique statistique, combinatoire

Plateforme et Master

- cluster de calcul (Christophe Cérin)
- master informatique (Younès Bennani)

Implication dans des structures

Fédération de Recherche MathSTIC, FR3734 (Christophe Fouqueré)

- 3 laboratoires :
  - LAGA, UMR 7539 (Philippe Souplet)
  - LIPN, UMR 7030 (Laure Petrucci)
  - L2TI, EA 3043 (Azeddine Beghdadi)
- 3 axes principaux :
  1. Optimisation et apprentissage appliqués aux contenus numériques
  2. Calcul haute-performance, systèmes distribués
  3. Physique mathématique, physique statistique, combinatoire

Plateforme et Master

- cluster de calcul (Christophe Cérin)
- master informatique (Younès Bennani)

Equipe A3 - Sciences des données

Yann Chevaleyre
resp. de l'équipe Apprentissage Artificiel
Laboratoire LIPN - Univ. Paris 13

Apprentissage non-supervisé

Transfert par
Factorisation non négative
[Y. Bennani, Y. Redko]

Apprentissage
non supervisé massif
[M. Lebbah, A. Azzag]

Problème du transfert

Données d'un certain type

Peuvent-elles aider pour apprendre sur d'autres données

Factorisation non négative

$$X \simeq WH^T, X \in \mathbb{R}^{n \times d}, W \in \mathbb{R}^{n \times k}, H \in \mathbb{R}^{d \times k}$$

$$X, W, H \geq 0.$$

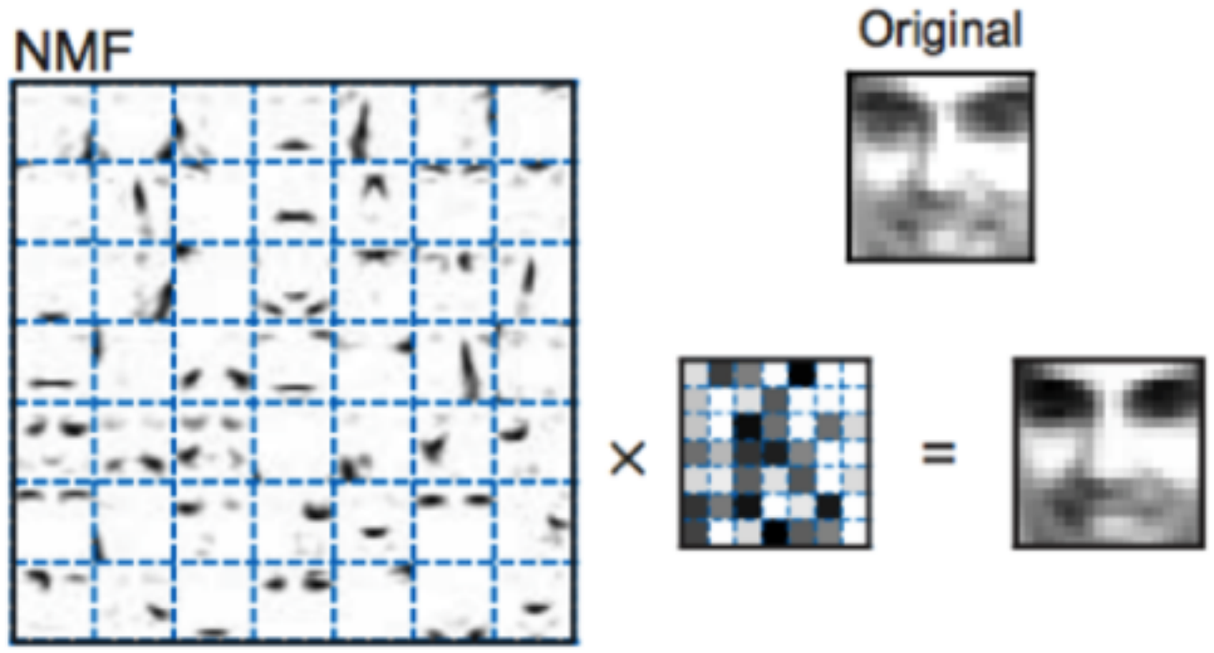
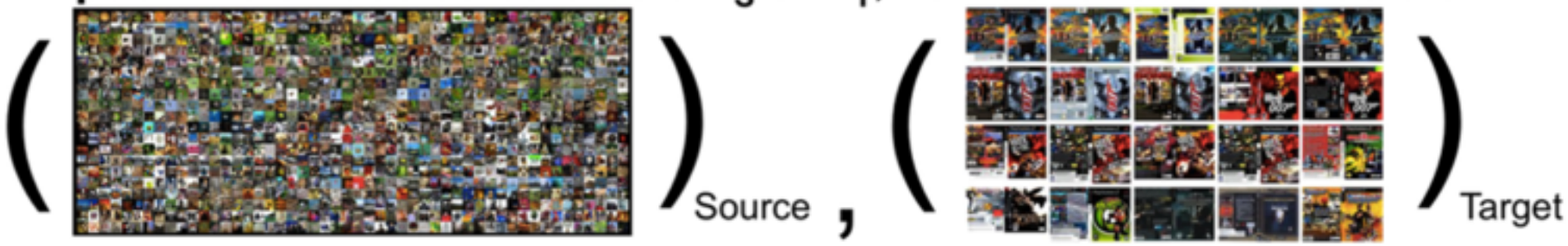


Figure: Figure depicted from (Lee and Seung, 1999)

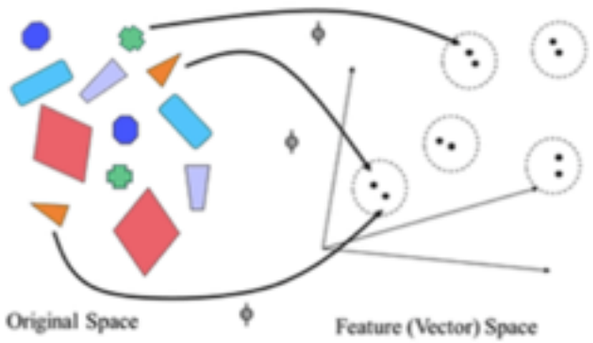
Input: Two unlabeled data sets $\mathbf{X_S}$ et $\mathbf{X_T}$, number of desired clusters **k**



Initialization: Compute Gram matrices $\mathbf{K_S}$ et $\mathbf{K_T}$ and $\hat{\mathbf{A}}(\mathbf{K_S},\mathbf{K_T})$

$$K : R^n \rightarrow H$$

H: Hilbert space – feature space



Using any arbitrary kernel function, i.e.:

$$k(x_i, x_j) = e^{\frac{||x_i - x_j||_2^2}{2\sigma^2}}$$

Kernel alignment optimization

$$\max \langle K_S, K_{ST} \rangle_F$$

$$K_{ST} = \sum_{n=1}^{k} \alpha_n K_n(x_{t_i}, x_{t_j})$$

$$\forall n, \alpha_n \geq 0.$$

Kernel NMF decomposition

$$K_{ST} = W_{ST} H_{ST}^T$$

Find $\mathbf{W_{ST}}$ such that:

$$W_{ST}^* = \arg\min_{W_{ST}} DBI(K_T).$$

Use $\mathbf{W_{ST}}$ as a bridge for transfer:

$$X_T = W_{ST}^* H_T^T$$

Date

Evaluation

We use the Office[Saenko et al., 2010]/Caltech[Gopalan et al., 2011] data set which consists of four domains:

- Amazon (**A**) - images from online merchants (958 images with 800 features from 10 classes);
- Webcam (**W**) - set of low-quality images by a web camera (295 images with 800 features from 10 classes);
- DSLR (**D**) - high-quality images by a digital SLR camera (157 images with 800 features from 10 classes);
- Caltech (**C**) - well-known data set for object recognition (1123 images with 800 features from 10 classes).

D

Evaluation

Table: Purity values on Office/Caltech data set obtained using BC-NMF

Domain pair (Source → Target)	**C-NMF**	**Kernel alone**	**TSC**	**BC-NMF**
C → A	33.24	40.34	43.32	**64.88**
C → W	46.78	56.00	52.54	**60.69**
C → D	46.5	47.33	54.14	**81.33**
A → C	24.89	35.33	46.03	**59.29**
A → W	46.78	56.00	53.22	**60.69**
A → D	46.5	47.33	51.59	**76.0**
W → C	24.89	35.33	**62.71**	58.97
W → A	33.24	40.34	61.36	**77.93**
W → D	46.5	47.33	59.66	**76.0**
D → C	24.89	35.33	**54.14**	52.0
D → A	33.24	40.34	54.78	**78.0**
D → W	46.78	56.00	55.59	**70.0**

Apprentissage non-supervisé

Transfert par
Factorisation non négative
[Y. Bennani, Y. Redko]

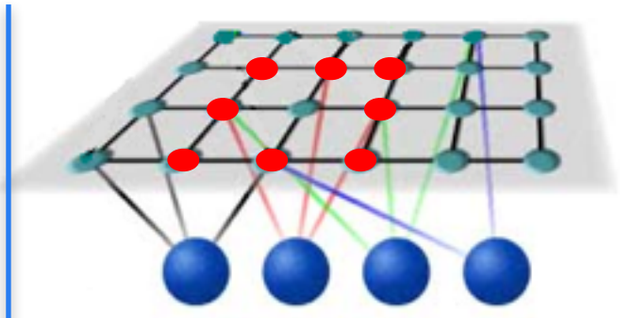
Apprentissage
non supervisé massif
[M. Lebbah, A. Azzag]

APPRENTISSAGE NON SUPERVISÉ (M. LEBBAH, H. AZZAG)

Continues

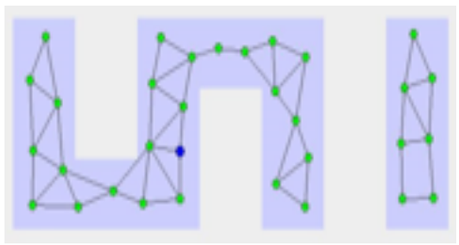
Binaires,catégorielles, mixtes

Data stream

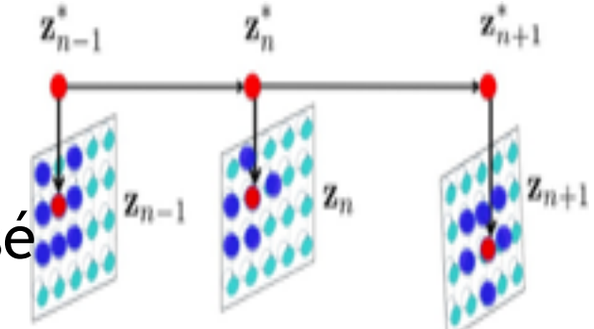


Visualisation

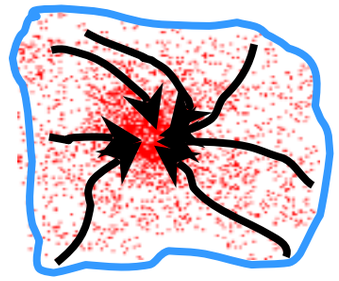
Clustering

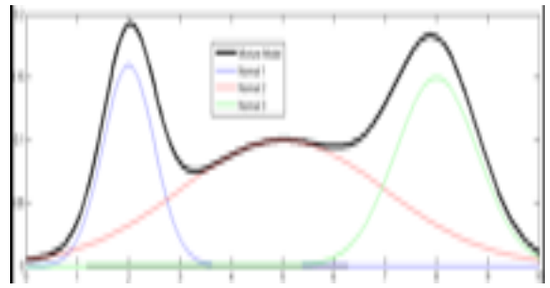


Topologie incrémentale



Modèle de Markov caché auto-organisé





$$p(\mathbf{x}) = \sum p(c)p(\mathbf{x}/c)$$

$$J^T_{som}(\mathcal{W},\phi) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c \in C}^{C} K^T(\delta(c,\phi(\mathbf{x}_i)))\|\mathbf{x}_i - \mathbf{w}_c\|^2$$

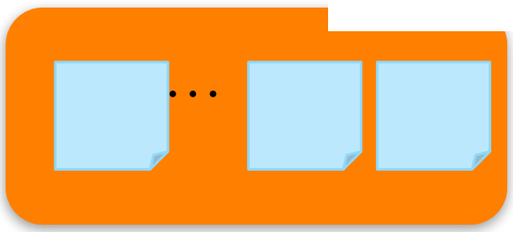
$$\mathcal{J}^T_{bin}(\phi,\mathcal{W},\Pi) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c \in C} K^T(\delta(c,\phi(\mathbf{x}_i)))\Pi^\beta\|\mathbf{x}_i - \mathbf{w}_c\|^2$$

$$\mathcal{R}_{wBiTM}(\mathbf{W},\mathbf{Z},\mathbf{G},\mathbf{\Pi}) = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{x^j_i \in B^l_k}\sum_{r=1}^{K}\mathcal{K}^T(\delta(r,k))(\pi^l_r x^j_i - g^l_r)^2$$

Big data et apprentissage Massivement distribué

(M. Lebbah, H. Azzag, D. Bouthinon)

Un projet Big Data (2013-2016)



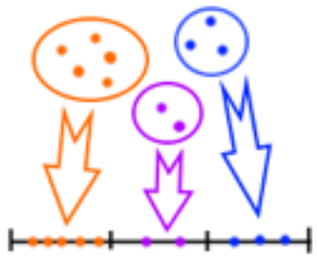
G-Stream

Visitez ce site
https://lipn.univ-paris13.fr/bigdata

Bientôt !
https://github.com/Spark-clustering-notebook/

Version Batch : NNMS

Nearest neighbour mean shift clustering (NNMS).



n dimensions

1 dimension

+

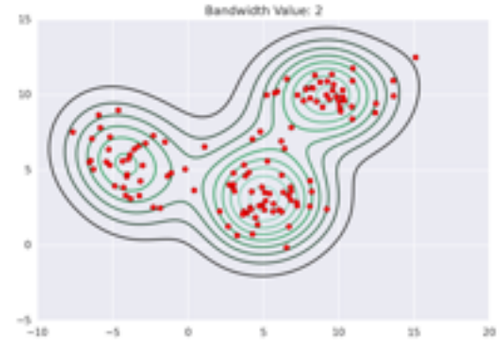


Image originale
481x321

Image à 43 clusters
200 blocs
20 min (4 cœurs)
4 min (20 cœurs)

La même qualité d'image est obtenue en 8h de calcul sous R

Conclusion / Défis

- Nombreux liens avec industriels / laboratoires
- Données hétérogènes - prise en compte du temps, les modalités
- Visualisation / réduction de dimensions
- Algorithmes en ligne et parallèles