

Apprentissage et classification par méthodes collaboratives

Comment choisir ses collaborateurs et qu'échanger avec eux ?

Antoine Cornuéjols

AgroParisTech – INRA MIA 518

Équipe LINK

Motivation

- « *The wisdom of crowd* »

[James Surowiecki, 2004]

- Estimation du **poids d'un panier** dans un marché
787 participants

[Francis Galton¹, 1906 (85 ans)]



¹ anthropologue, explorateur, géographe, inventeur, météorologue, proto-généticien, psychométricien et statisticien

Motivation

- « *The wisdom of crowd* »

[James Surowiecki, 2004]

- Estimation du **poids d'un panier** dans un marché

787 participants

- Le meilleur = plus d'un **centième** d'erreur
- Moyenne : moins d'un **millième** d'erreur

[Francis Galton¹, 1906 (85 ans)]



¹ anthropologue, explorateur, géographe, inventeur, météorologue, proto-généticien, psychométricien et statisticien

Pourquoi / comment ça marche

- « Experts faibles »

Estimations « bruitées »

- Non biaisées
- Symétriques
- Indépendantes



Combinaison simple :
la moyenne

Plan

1. Contexte et motivations
2. Méthodes collaboratives en IA
3. Méthodes collaboratives en Apprentissage Automatique
4. Quid du clustering ?
5. Bilan

Motivations (1)

- Améliorer les performances des solutions (clusterings) en les combinant

Méthodes « d'ensemble » non supervisées

Motivations (2)

- Énormément de données produites « localement »
 - Hôpitaux
 - Succursales d'entreprises
 - Internet des Objets
- Difficile de centraliser ces données -> traitements locaux
- Qui peuvent cependant bénéficier d'échanges

Les méthodes « collaboratives »

Combinaison de plusieurs résultats de clustering

1. Sur les mêmes données

- Des **algorithmes** différents (biais / initialisation)
- Des **mesures** différentes
- Clustering **multi-vues / multi-objectifs**
- Recherche de **consensus** : clustering « *coopératif* »

Combinaison de plusieurs résultats de clustering

2. Sur des données différentes : multi-sources

- Recherche d'un clustering global
- Recherche de solutions locales : clustering « collaboratif »
 - Sous contraintes d'échanges limitées avec les autres « experts »
 - ✓ Vie privée
 - ✓ Sécurité sur les échanges

Combinaison de plusieurs résultats de clustering

1. Sur les mêmes données

- Des **algorithmes** différents (biais / initialisation)
- Des **mesures** différentes
- Clustering **multi-vues / multi-objectifs**
- Recherche de **consensus** : clustering « *coopératif* »

2. Sur des **données différentes** : multi-sources

- Recherche d'un **clustering global**
- Recherche de **solutions locales** : clustering « *collaboratif* »
 - Sous contraintes d'échanges limitées avec les autres « experts »
 - ✓ Vie privée
 - ✓ Sécurité sur les échanges

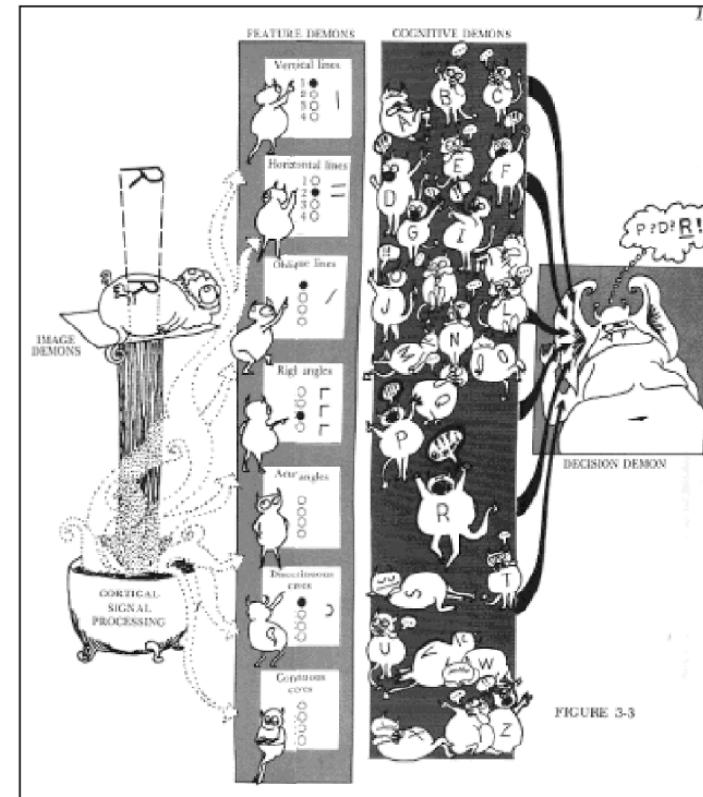
Plan

1. Contexte et motivations
2. Méthodes collaboratives en IA
3. Méthodes collaboratives en Apprentissage Automatique
4. Quid du clustering ?
5. Bilan

Méthodes collaboratives en Intelligence Artificielle

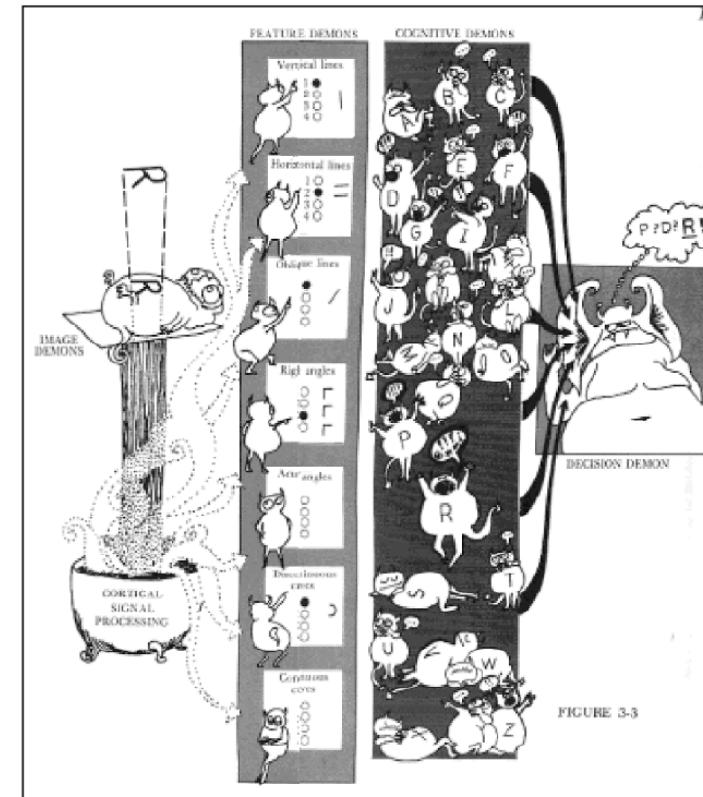
Pandemonium

- First Pandemonium (1958)
 - Oliver Selfridge « *Pandemonium: A Paradigm for Learning* »



Pandemonium

- First Pandemonium (1958)
 - Oliver Selfridge « *Pandemonium: A Paradigm for Learning* »
 - A hierarchical architecture of « demons » to solve problems + a suggestion for a learning mechanism
 - « **Data** demons : specialized in some types of input data (horizontal line, circle subparts, ...)
 - « **Cognitive** demons »: integrate information coming from sub-levels demons
 - « **Decision** demons » : make decisions about the interpretation



Demons shout with a **strength** in proportion to their **certainty** in their claim

This « strength » is set through **learning**

Hearsay II (1975)

- Speech recognition
 - The DARPA Speech Understanding Research (SUR) program

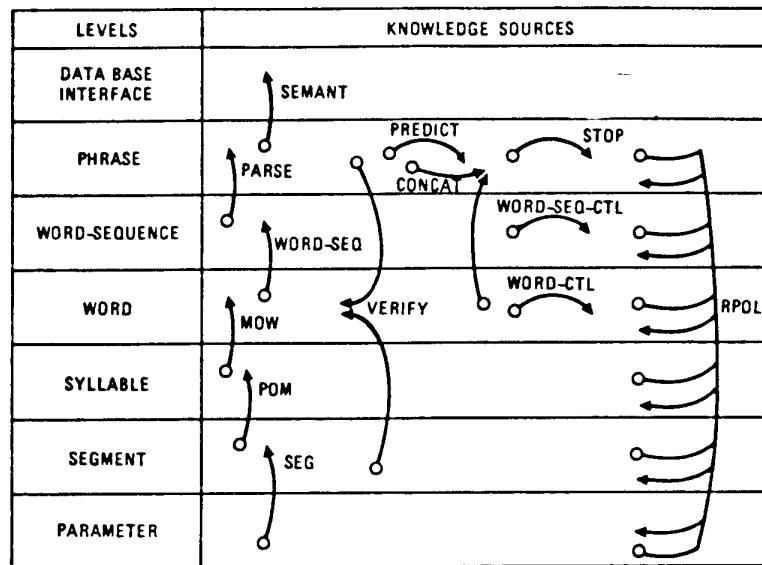


FIGURE 2. The levels and knowledge sources of September 1976. KSs are indicated by vertical arcs with the circled ends indicating the input level and the pointed ends indicating output level.

Signal Acquisition, Parameter Extraction, Segmentation, and Labeling:

- SEG: Digitizes the signal, measures parameters, and produces a labeled segmentation.

Word Spotting:

- POM: Creates syllable-class hypotheses from segments.
- MOW: Creates word hypotheses from syllable classes.
- WORD-CTL: Controls the number of word hypotheses that MOW creates.

Phrase Island Generation:

- WORD-SEQ: Creates word-sequence hypotheses that represent potential phrases from word hypotheses and weak grammatical knowledge.
- WORD-SEQ-CTL: Controls the number of hypotheses that WORD-SEQ creates.
- PARSE: Attempts to parse a word sequence and, if successful, creates a phrase hypothesis from it.

Phrase Extending:

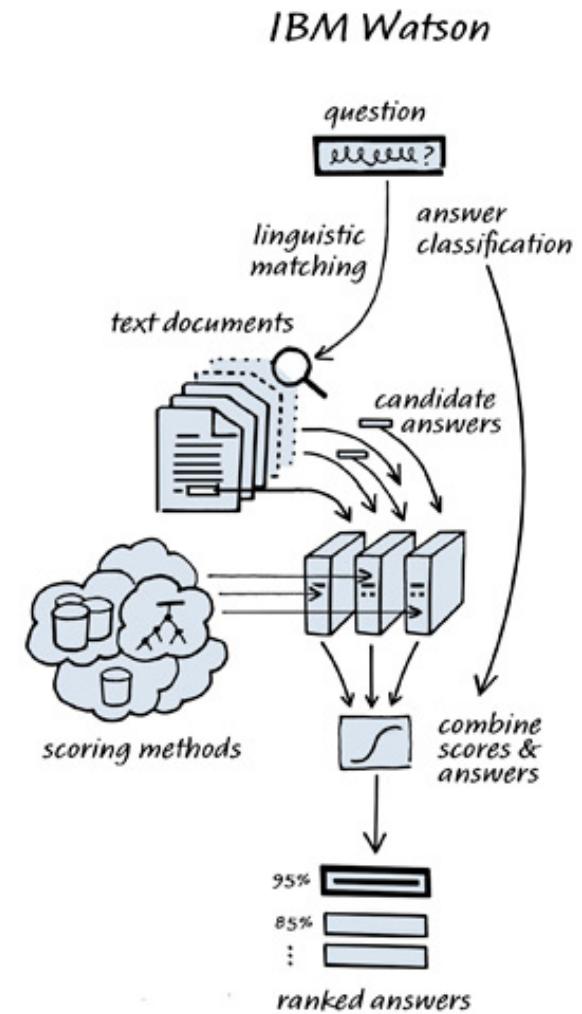
- PREDICT: Predicts all possible words that might syntactically precede or follow a given phrase.
- VERIFY: Rates the consistency between segment hypotheses and a contiguous word-phrase pair.
- CONCAT: Creates a phrase hypothesis from a verified contiguous word-phrase pair.

Rating, Halting, and Interpretation:

- RPOL: Rates the credibility of each new or modified hypothesis, using information placed on the hypothesis by other KSs.
- STOP: Decides to halt processing (detects a complete sentence with a sufficiently high rating, or notes the system has exhausted its available resources) and selects the best phrase hypothesis or set of complementary phrase hypotheses as the output.
- SEMANT: Generates an unambiguous interpretation for the information-retrieval system which the user has queried.

Plus récemment : WATSON

- Search the best answer to open questions
- Require a the exploration of a huge search space
 - Documents
 - Internet



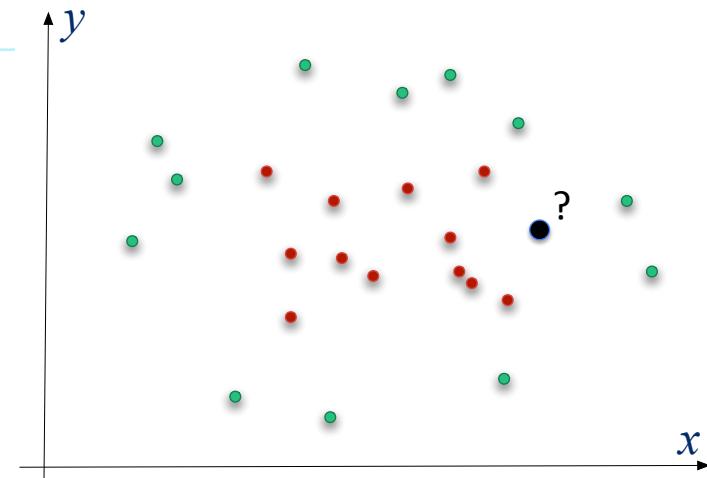
Co-construction d'une **interprétation**

- Les experts sont « **donnés** »
- Les experts **interviennent de manière opportuniste ou à tour de rôle**
- interagissent par **échanges de contraintes**

Co-learning et co-clustering

Les k plus-proches-voisins

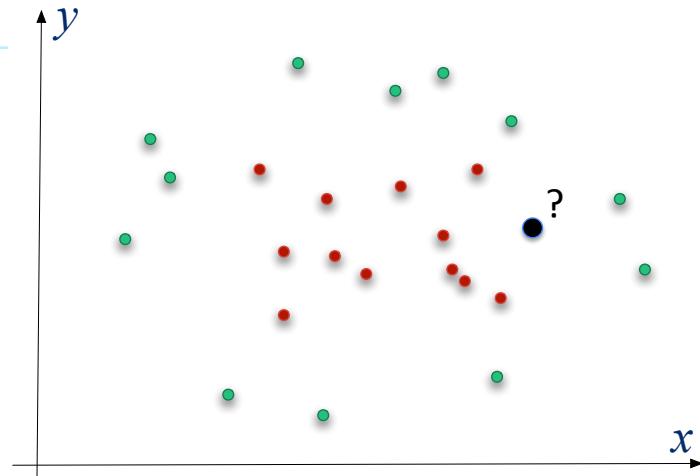
Sélection des collaborateurs



- Les « experts » (voisins) sont sollicités en fonction de la question
 - Approche « **lazy learning** » ou **transductive** learning

Les k plus-proches-voisins

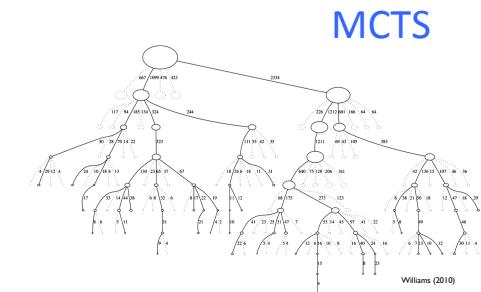
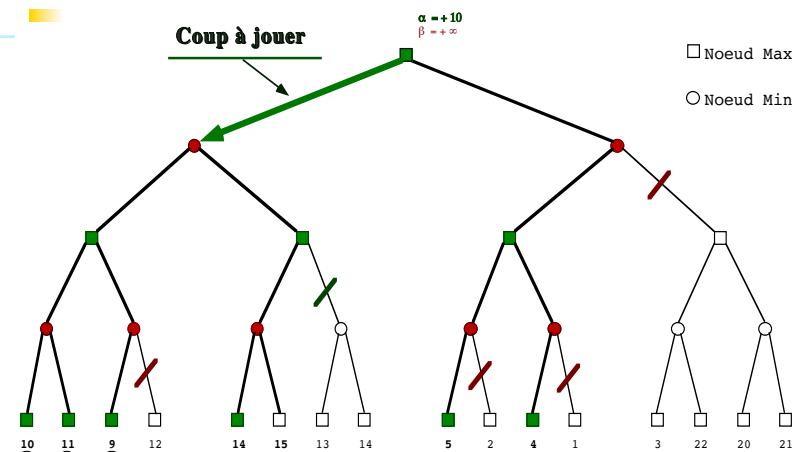
Sélection des collaborateurs



- Les « experts » (voisins) sont sollicités en fonction de la question
 - Approche « **lazy learning** » ou **transductive** learning
- Questions
 - Comment choisir ses « voisins » ? Quelle **distance** ?
 - Comment **combiner** les réponses : (e.g. **vote majoritaire (pondéré)**)

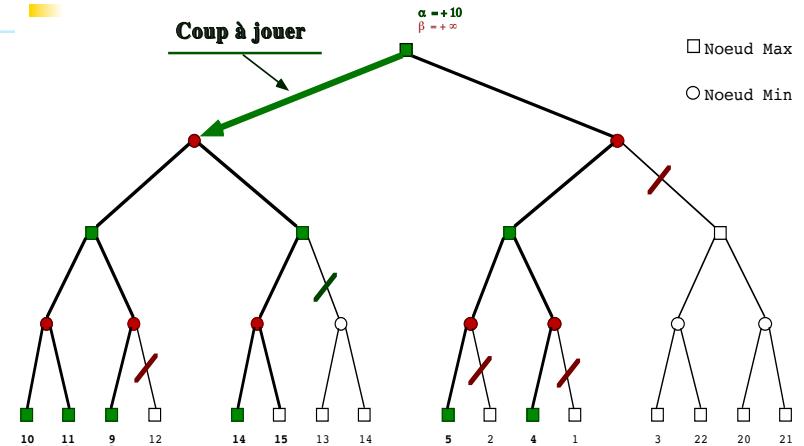
Algorithmes de jeu

Construction des collaborateurs



Algorithmes de jeu

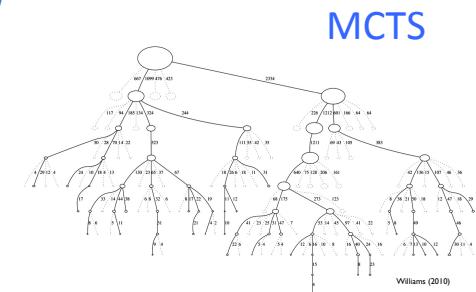
Construction des collaborateurs



- Quel coup jouer ?

Fonction d'évaluation imprécise à la racine (situation courante)

1. Sollicitation d'experts ayant une évaluation plus précise sur des cas possibles liés à la situation courante
2. Combinaison hiérarchique des évaluations par MinMax



Les experts peuvent travailler dans des espaces différents

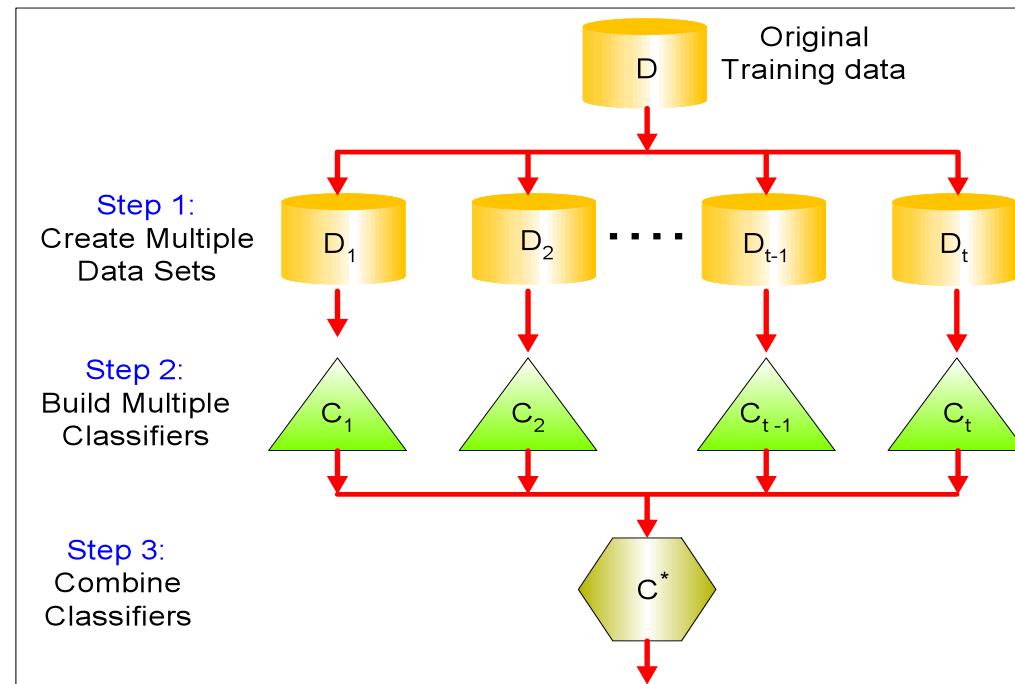
Plan

1. Contexte et motivations
2. Méthodes collaboratives en IA
3. Méthodes collaboratives en Apprentissage Automatique
4. Quid du clustering ?
5. Bilan

Méthodes collaboratives en Apprentissage Automatique

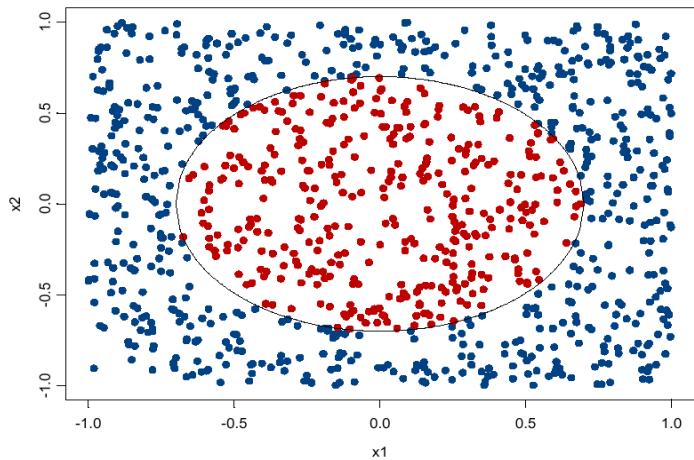
Le bagging et les « random forests »

- Construction de **solutions faibles** par apprentissage sur des sous-échantillons de données

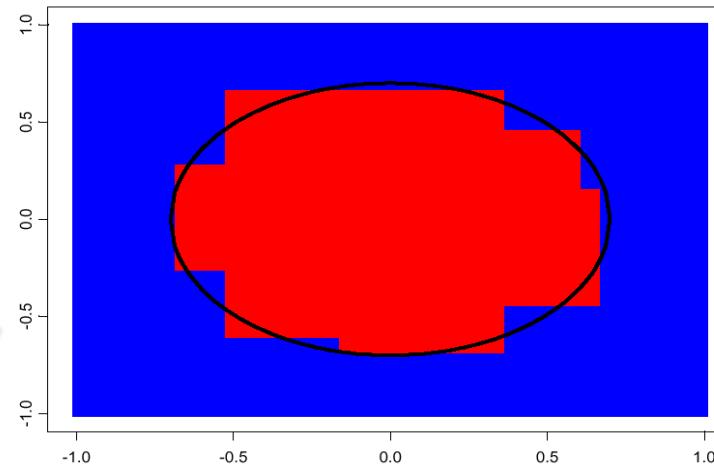


Le bagging et les « random forests »

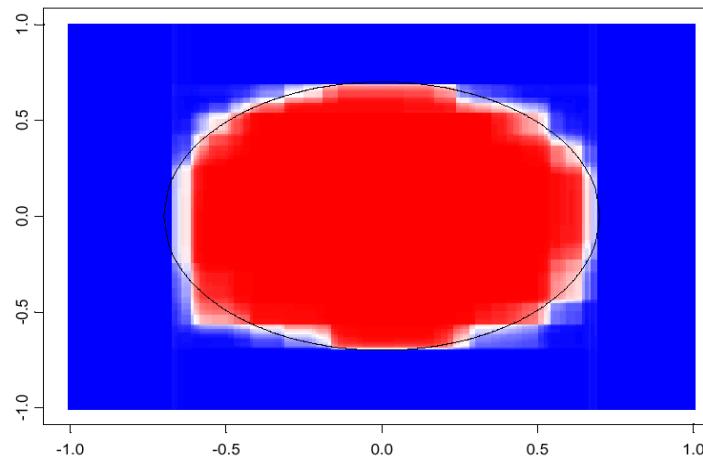
- Illustration



CART

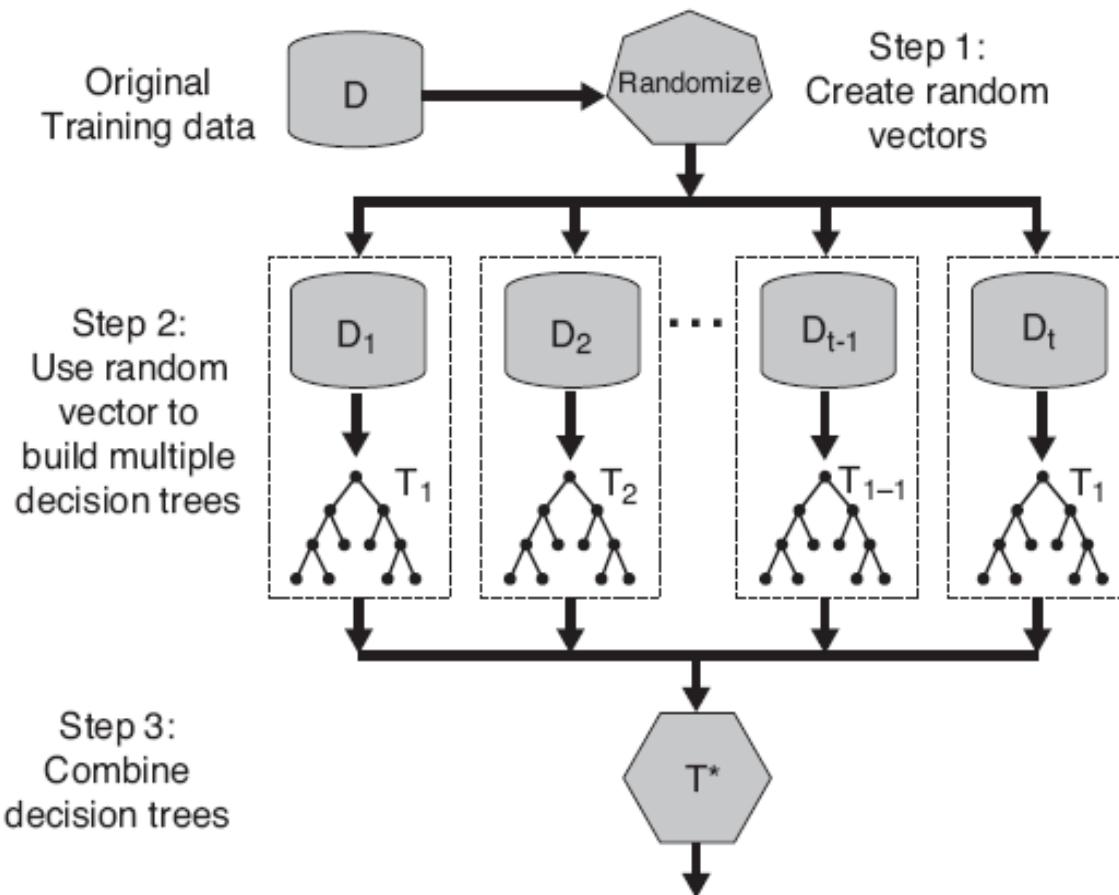


Bagging



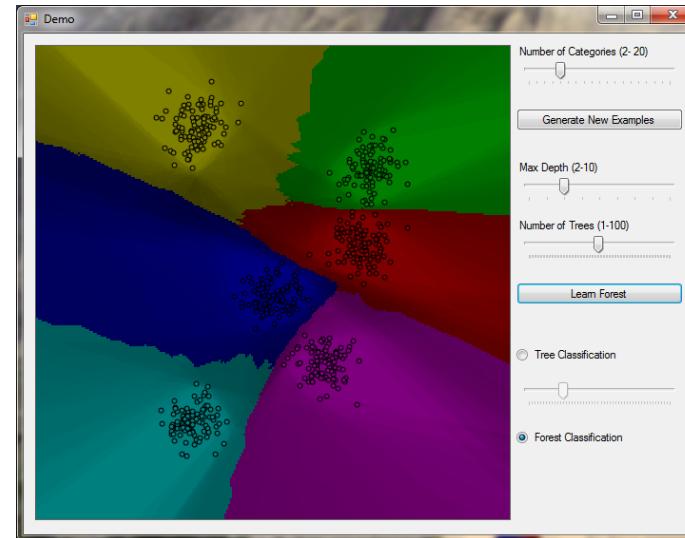
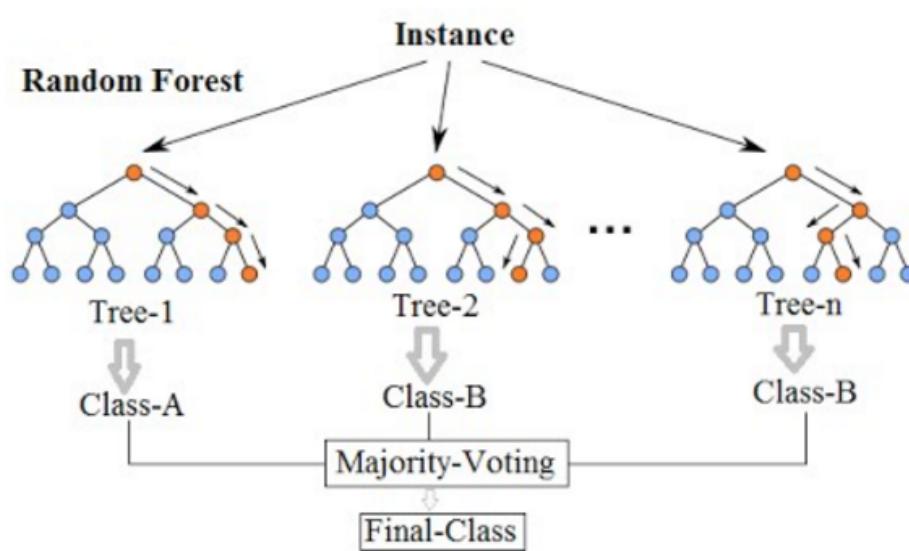
shades of blue/red indicate strength of vote for particular classification

Le bagging et les « random forests »



Le bagging et les « random forests »

- Random forests



Averaging together many trees in a forest can result in decision boundaries that look very sensible, and are even quite close to the max margin classifier.
(Shading represents entropy – darker is higher entropy).

Les méthodes d'ensemble

- La « sagesse » de la foule : combiner des méthodes / résultats
 - Limites d'une solution « directe »
 - **Biais incertain**
 - **Exploration** de l'espace des solutions sujettes à minima locaux
- On espère **additionner les forces** des méthodes faibles
Et compenser / **annuler les faiblesses**

Les méthodes d'ensemble

- La « sagesse » de la foule : combiner des méthodes / résultats
 - Limites d'une solution « directe »
 - Biais incertain
 - Exploration de l'espace des solutions sujettes à minima locaux
 - On espère additionner les forces des méthodes faibles
Et compenser / annuler les faiblesses
- Approches :
 - Connaissances complémentaires : *architecture « blackboard »*
 - Création d'experts indépendants : *le bagging / Random Forests*
 - Création d'experts complémentaires : *le boosting*

Les méthodes d'ensemble

- La « sagesse » de la foule : combiner des méthodes / résultats
 - Limites d'une solution « directe »
 - Biais incertain
 - Exploration de l'espace des solutions sujettes à minima locaux
 - On espère additionner les forces des méthodes faibles
 - Et compenser / annuler les faiblesses
- Approches :
 - Connaissances complémentaires : *architecture « blackboard »*
 - Création d'experts indépendants : *le bagging / Random Forests*
 - Création d'experts complémentaires : *le boosting*
- Questions :
 - Comment construire des experts faibles ?
 - Comment combiner les solutions faibles

Les approches collaboratives

- Motivation : pas assez d'informations pour résoudre la question « localement »

Les approches collaboratives

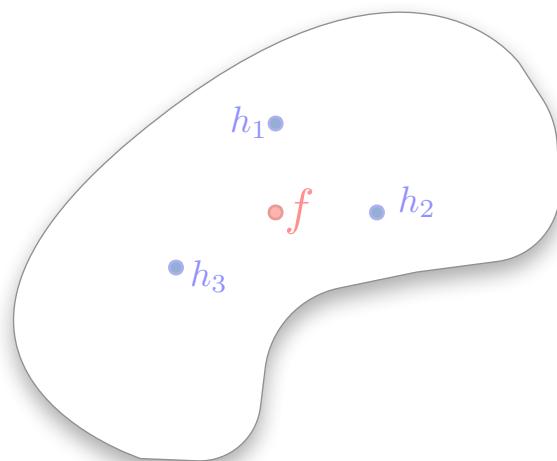
- Motivation : pas assez d'informations pour résoudre la question « localement »
- Approche : recours à des sources d'information externes
 - kNN : CBR
 - Analogie / Transfert

Les approches collaboratives

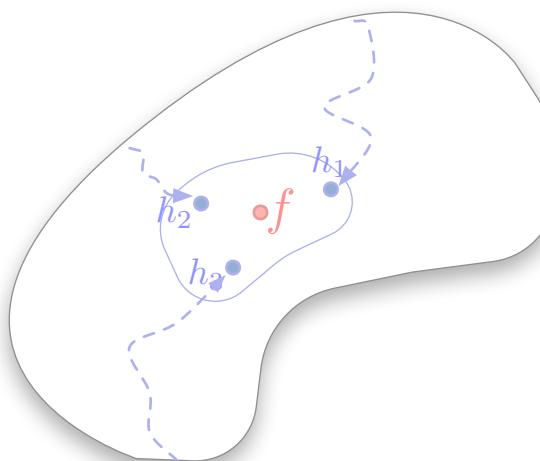
- Motivation : pas assez d'informations pour résoudre la question « localement »
- Approche : recours à des sources d'information externes
 - kNN : CBR
 - Analogie / Transfert
- Questions :
 - Choix des sources externes
 - Quelle information transmettre (et traduire) ?
 - Comment combiner les informations externes ?

Méthodes d'ensemble supervisées

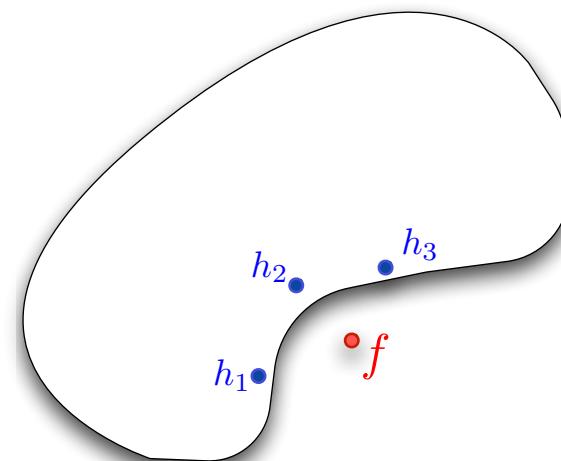
Justifications intuitives des méthodes d'ensemble



Justification
statistique



Justification
computationnelle



Justification
représentationnelle

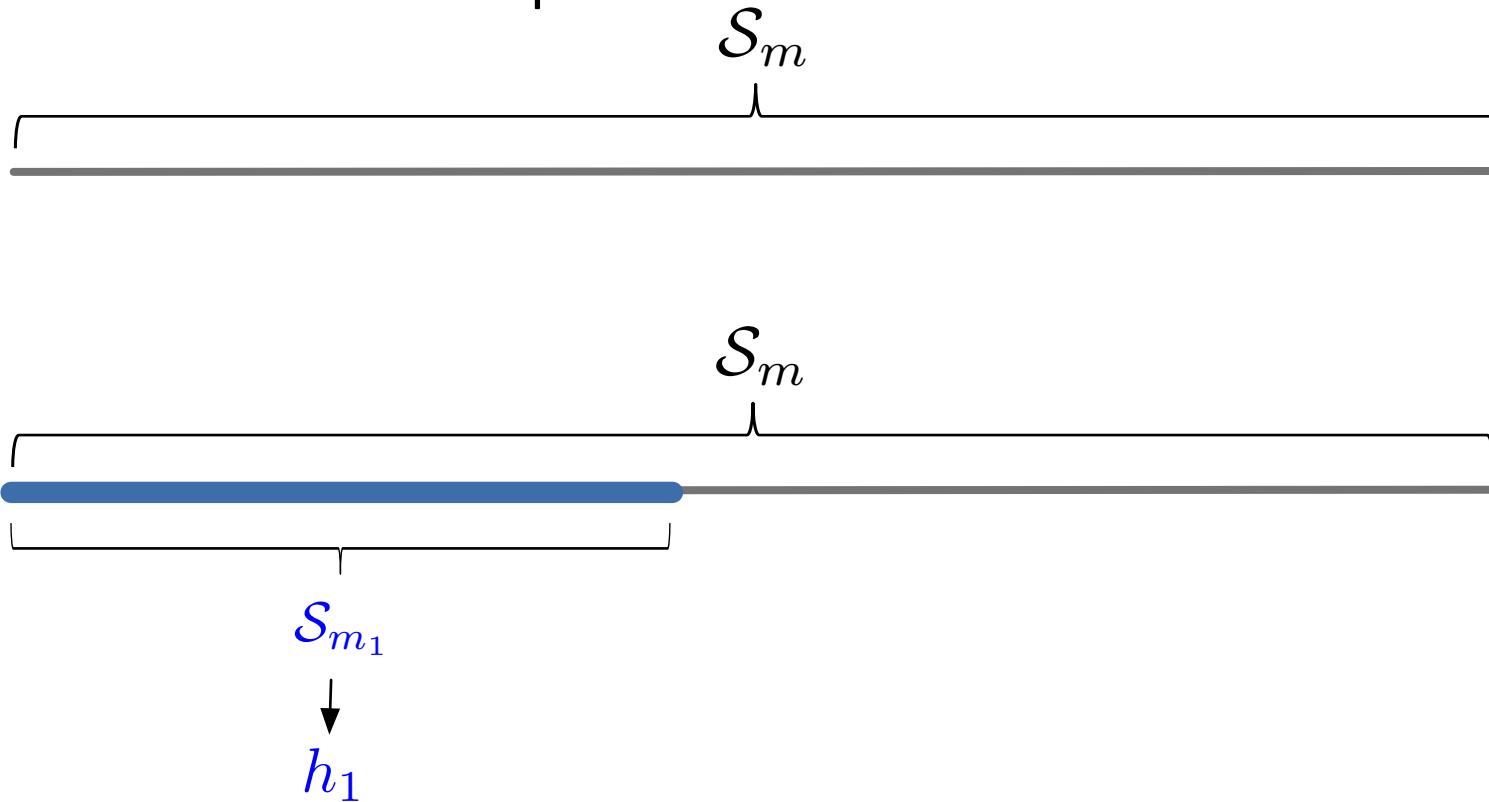
- Dietterich T. (2000) « Ensemble Methods in Machine Learning ». Proc. 1st Int. Workshop on *Multiple Classifier Systems*, Sardinia, Italy, 2000.

Une question théorique

- Apprentissage « **fort** » (PAC learning)
 - Une classe de fonctions \mathcal{F} est apprenable (au sens **fort**) si il existe un algorithme d'apprentissage \mathcal{A} qui pour toute distribution \mathcal{D}_X sur X , et pour toute fonction f est tel que :
$$\forall \varepsilon, \delta : \exists m(\varepsilon, \delta) \text{ st. } \text{Prob}[R(h_{\mathcal{S}}) > \varepsilon] \leq \delta$$
- Apprentissage « **faible** »
 - Une classe de fonctions \mathcal{F} est apprenable (au sens **faible**) si, pour $\gamma > 0$, il existe un algorithme d'apprentissage \mathcal{A} qui pour toute distribution \mathcal{D}_X sur X , et pour toute fonction f est tel que :
$$\forall \delta : \exists m(\delta) \text{ st. } \text{Prob}[R(h_{\mathcal{S}}) > 1/2 - \gamma] \leq \delta$$
- Sont-ils de nature différente ?

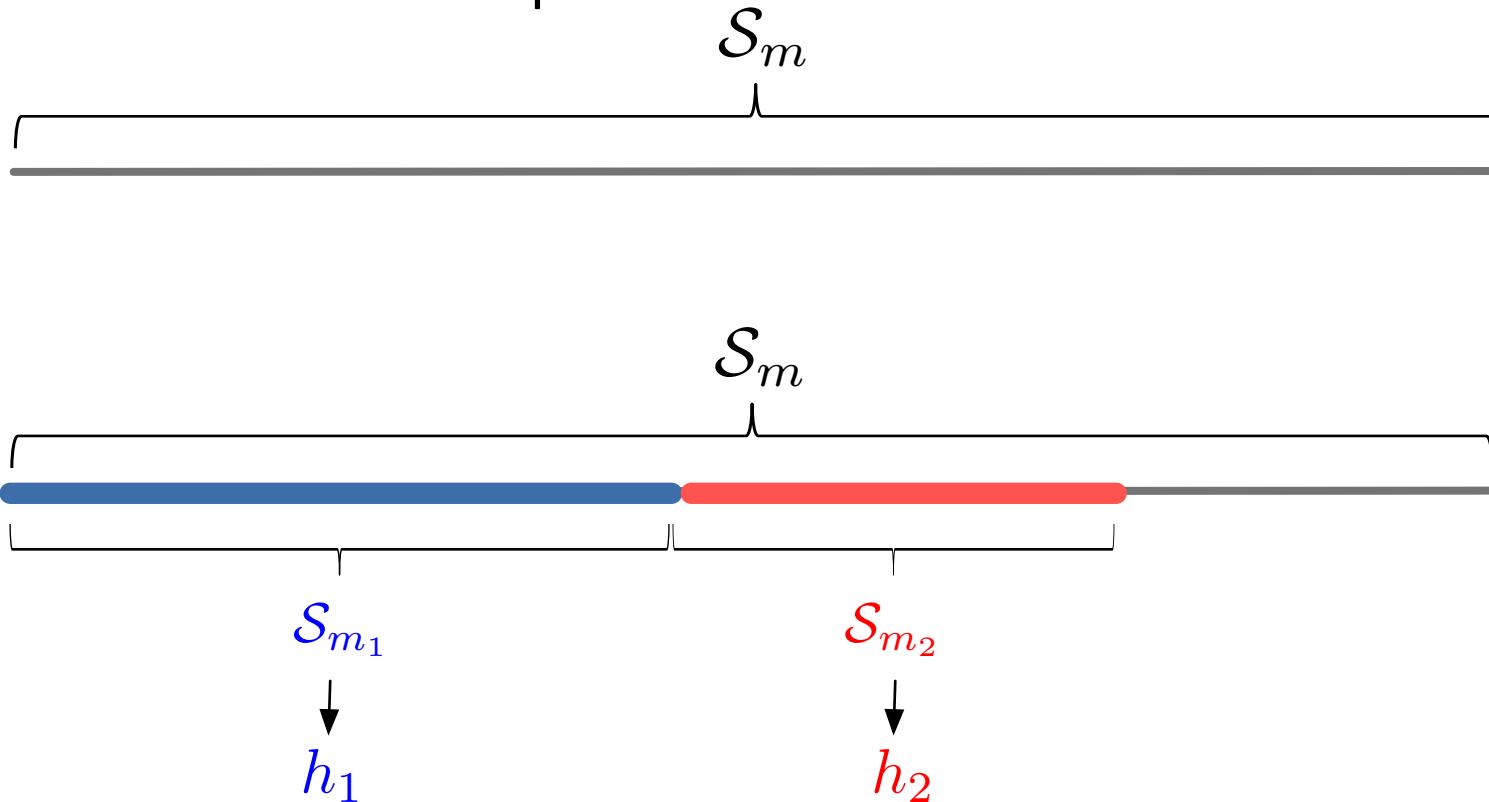
Comment engendrer les apprenants

- Une recette historique



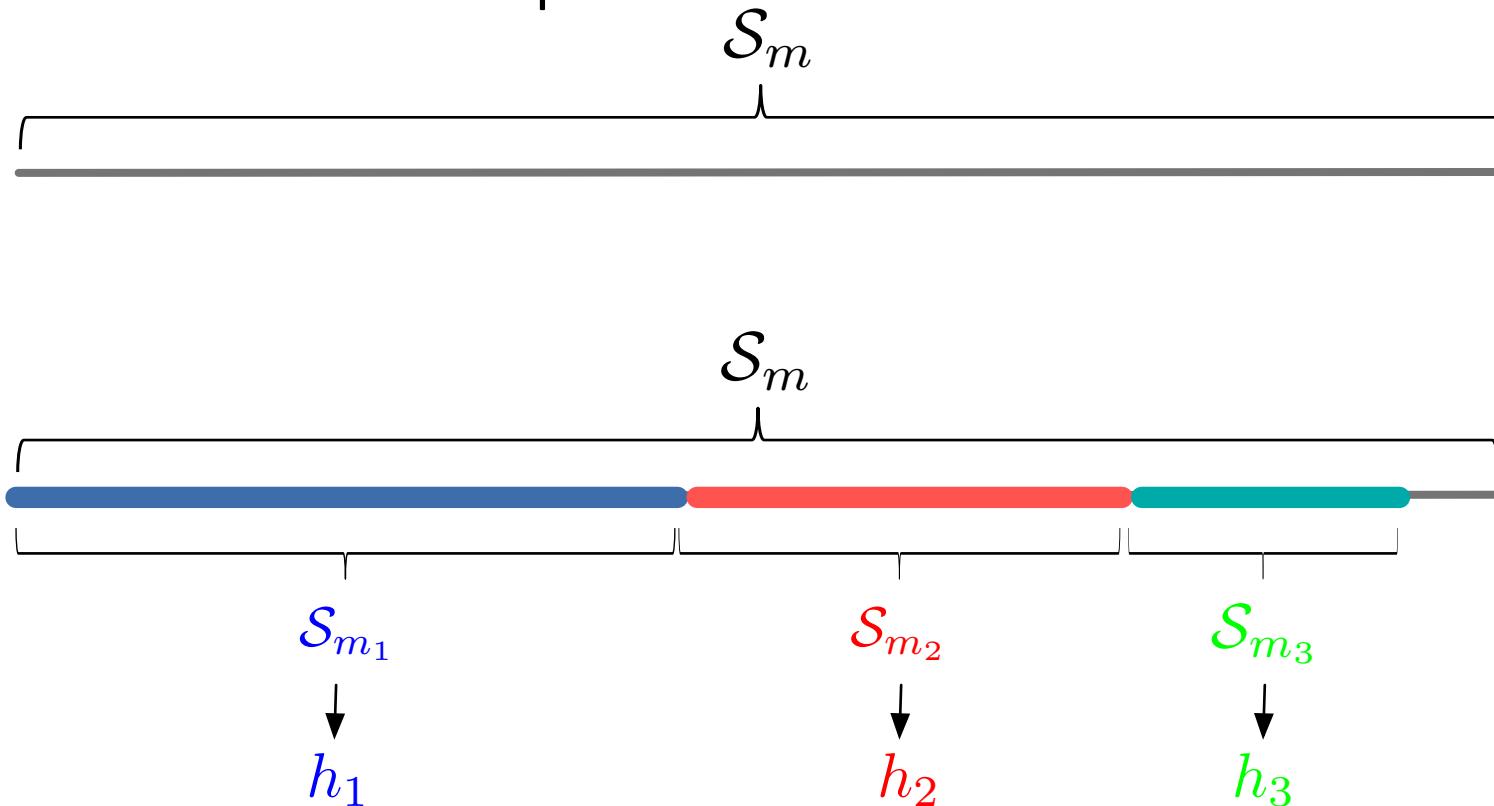
Comment engendrer les apprenants

- Une recette historique



Comment engendrer les apprenants

- Une recette historique



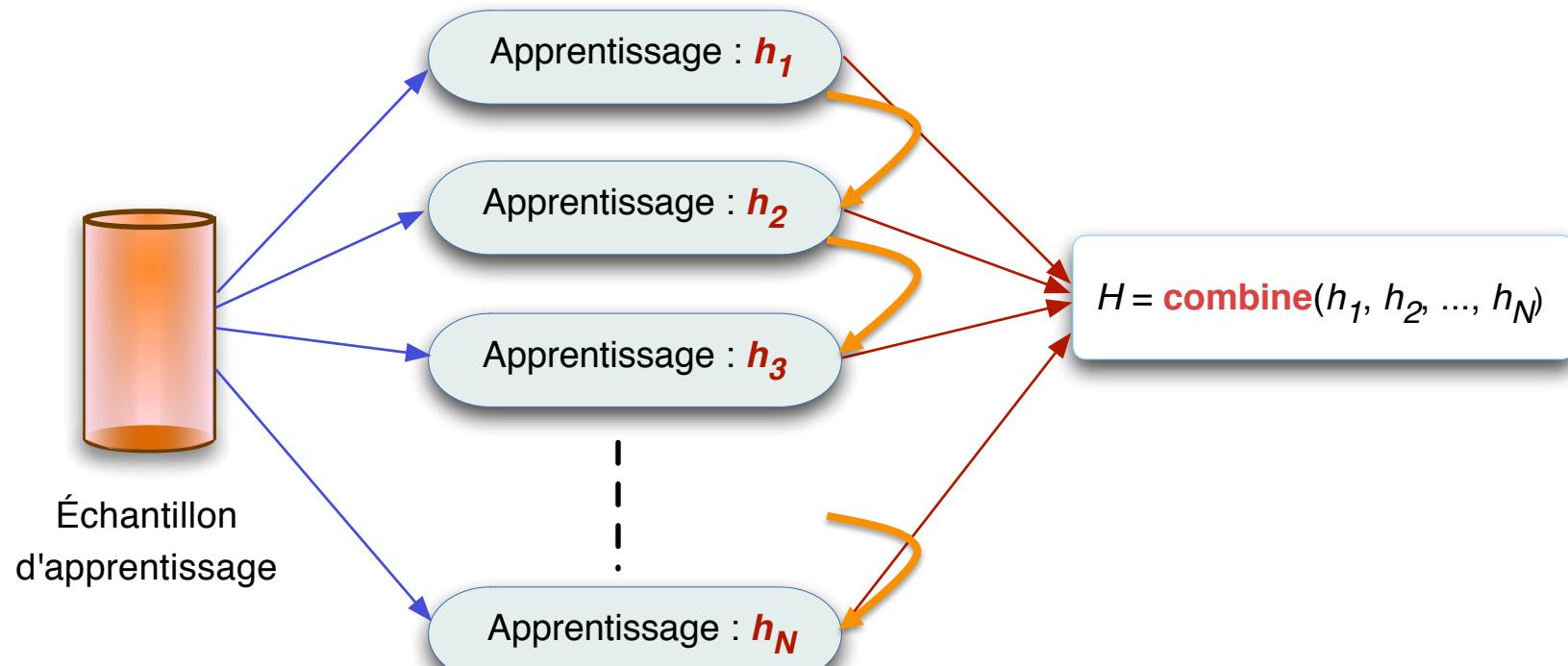
$$H(\mathbf{x}) = \text{sign}\left(h_1(\mathbf{x}) + h_2(\mathbf{x}) + h_3(\mathbf{x}) \right)$$

Comment engendrer les apprenants (suite)

- Modifier l'échantillon d'apprentissage à chaque étape
 - *En diminuant l'importance des exemples bien classés*
 - *En augmentant ----- mal -----*
 - De combien ?

Le boosting

- Schéma général



Boosting

- **boosting** = méthode générale pour convertir des règles de prédiction peu performantes en une règle de prédiction (très) performante

- Plus précisément :
 - Étant donné un algorithme d'apprentissage "faible" qui peut toujours retourner une hypothèse de taux d'erreur $\leq 1/2 - \gamma$

 - Un algorithme de boosting peut construire (de manière prouvée) une règle de décision (hypothèse) de taux d'erreur $\leq \varepsilon$

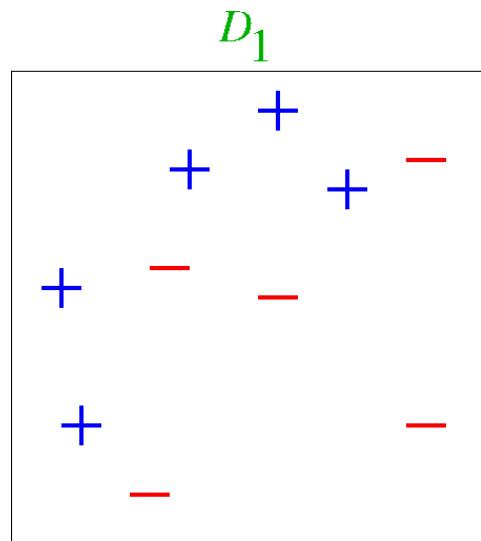
AdaBoost en plus gros

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right) > 0$$

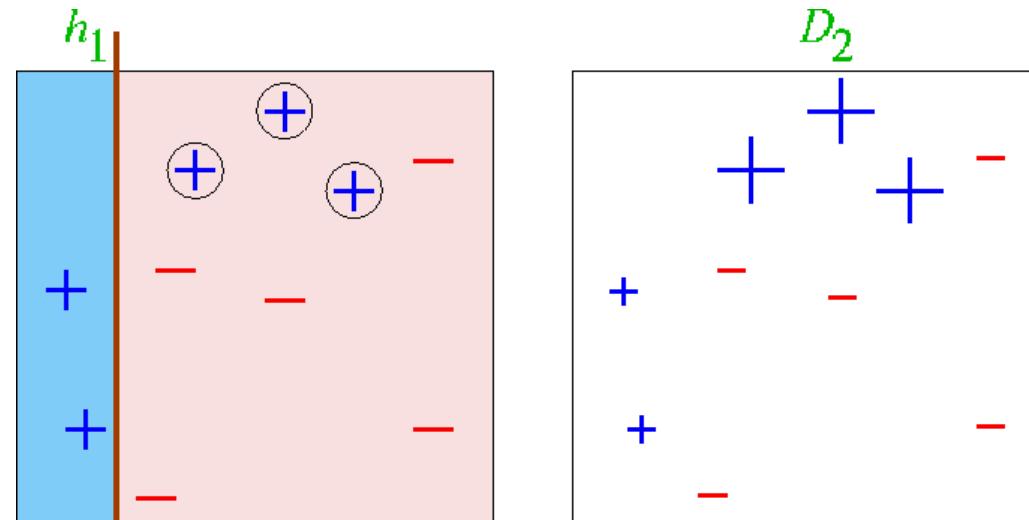
$$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$H_{\text{final}}(x) = \text{sgn} \left(\sum_t \alpha_t h_t(x) \right)$$

Exemple jouet



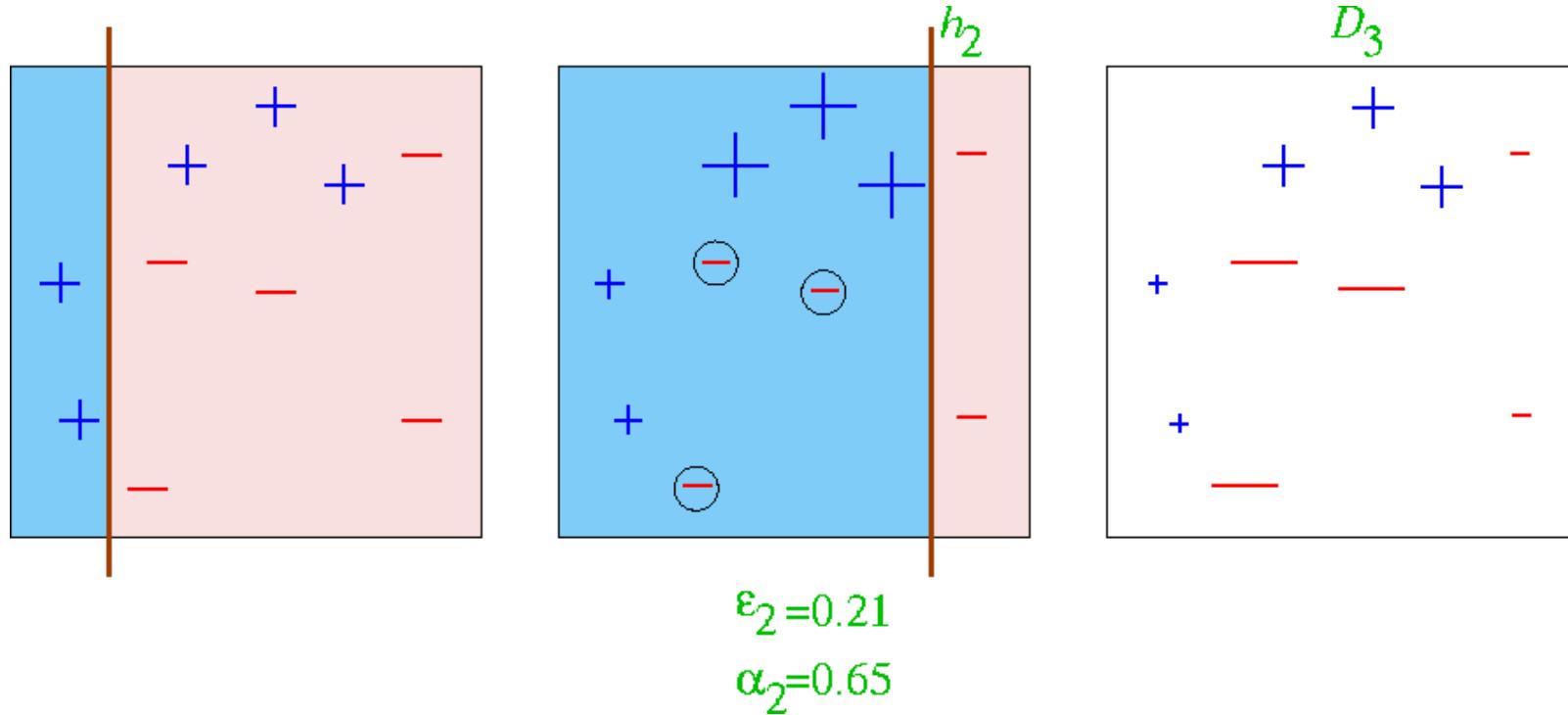
Étape 1



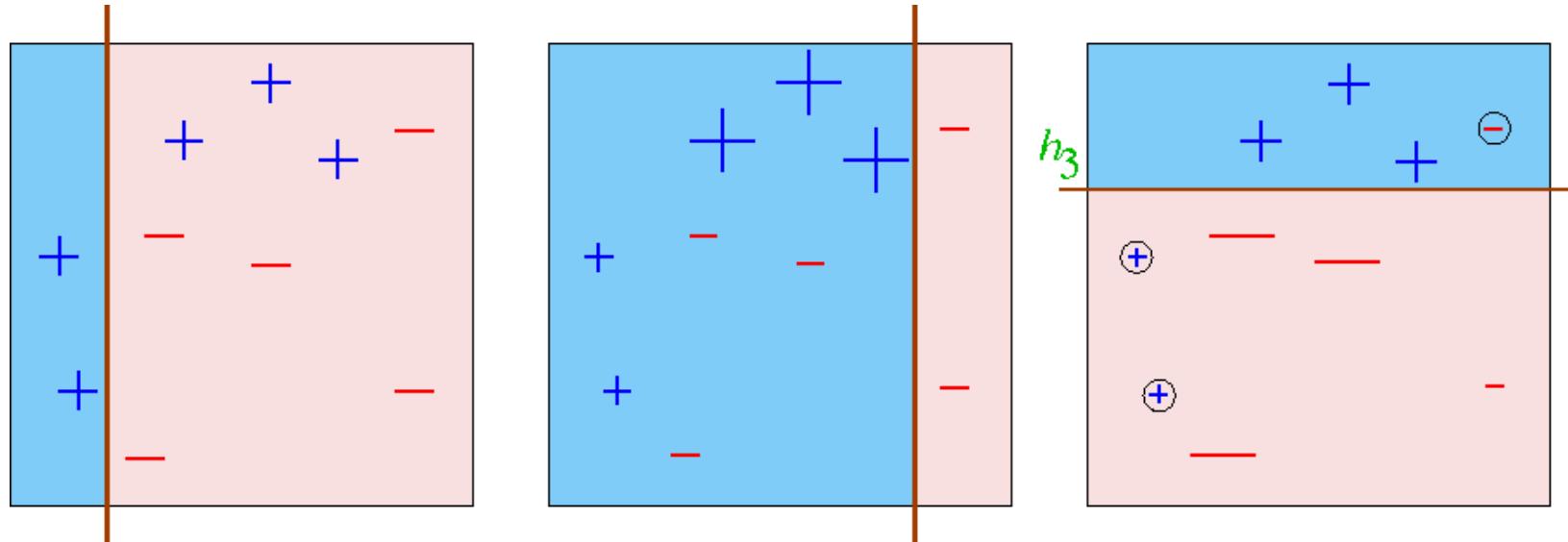
$$\epsilon_1 = 0.30$$

$$\alpha_1 = 0.42$$

Étape 2



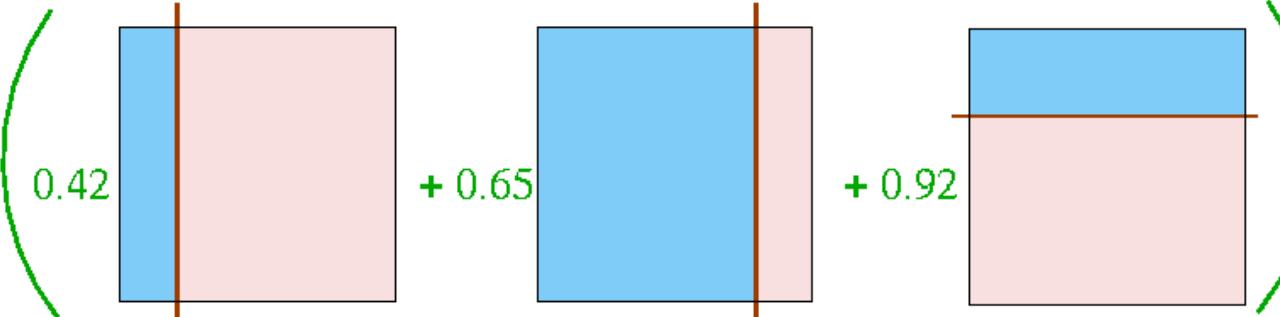
Étape 3

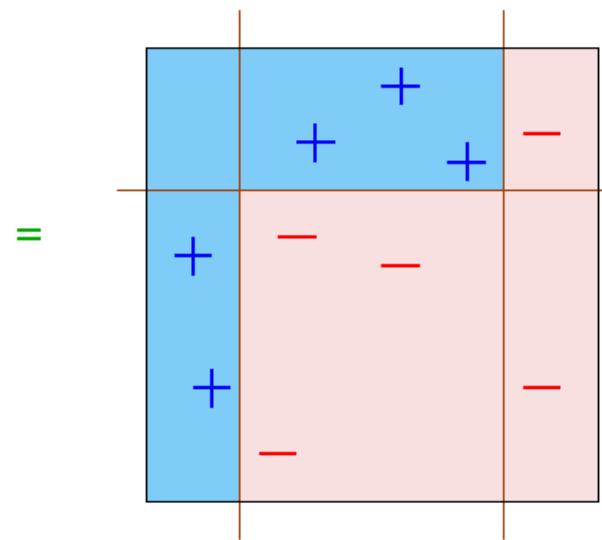


$$\varepsilon_3 = 0.14$$

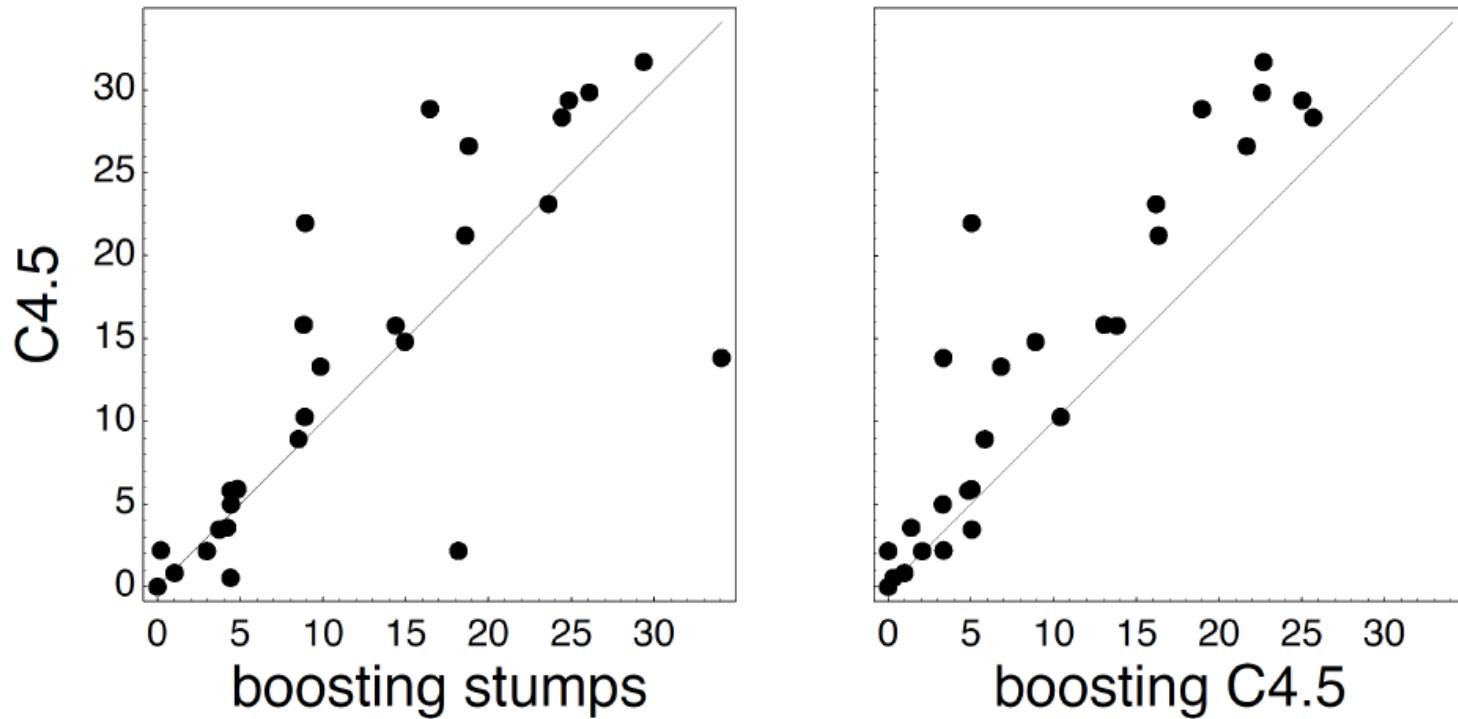
$$\alpha_3 = 0.92$$

Hypothèse finale

$$H_{\text{final}} = \text{sign} \left(0.42 + 0.65 + 0.92 \right)$$




Performances du boosting



Test error rate on 27 benchmark problems
x-axis: boosting; y-axis: base-line (C4.5)

Transduction

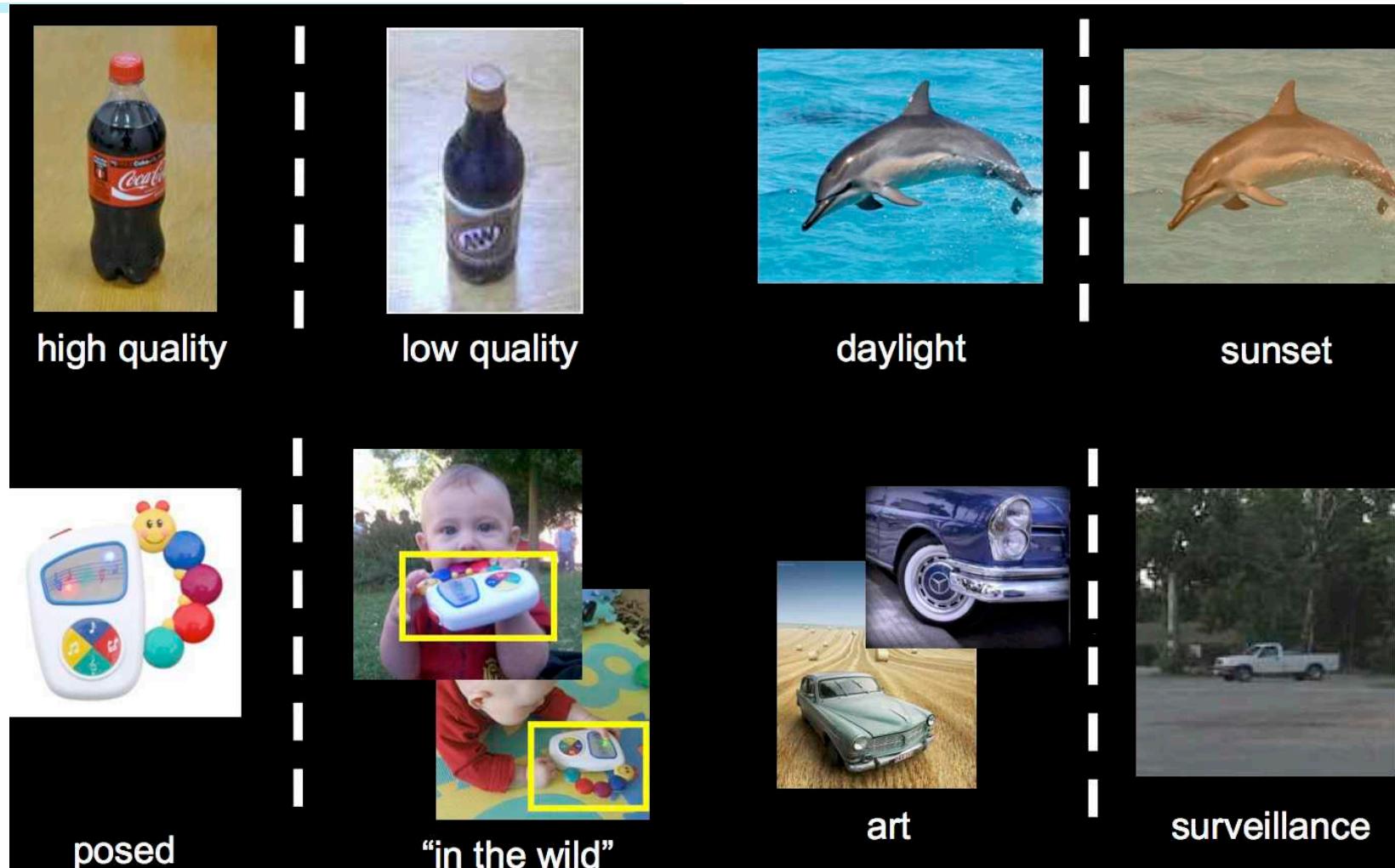
Transfert

+ méthode d'ensemble

=

Transboost

Examples: transfer learning in vision



[Xu, Saenko, Tsang “Domain Transfer” tutorial – CVPR’12]

Apprentissage par transfert

- Illustrations



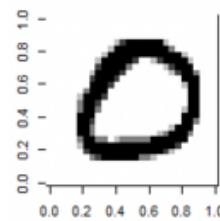
FIGURE 1: Trained model on the data source : is it a picture of a dog or a cat ?



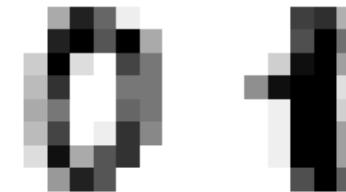
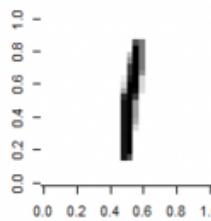
FIGURE 2: Model source transferred on the data target : is it a clip-art of a dog or a cat ?

Apprentissage par transfert

- Illustrations

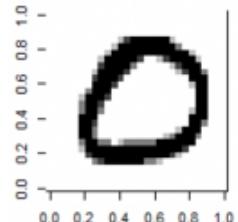


(a) Is it a zero or a one?

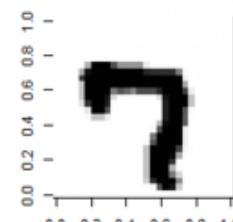
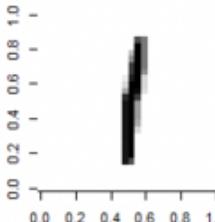


(b) Is it a zero or a one?

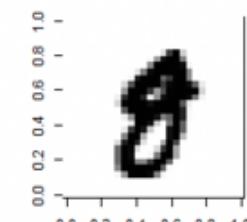
FIGURE 15: Transfer learning of the source model 0/1 mnist so that it can distinguish 0/1 sklearn digits



(a) Is it a zero or a one?

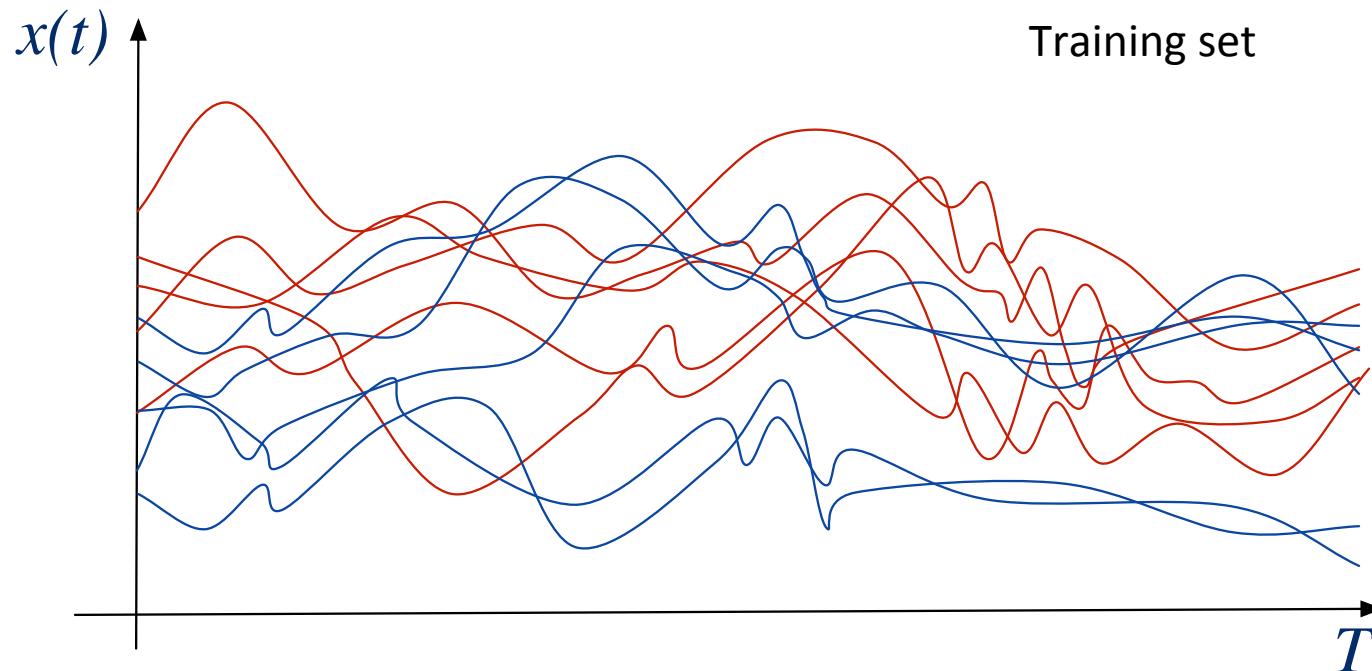


(b) Is it an eight or a seven?



La classification précoce de séries temporelles

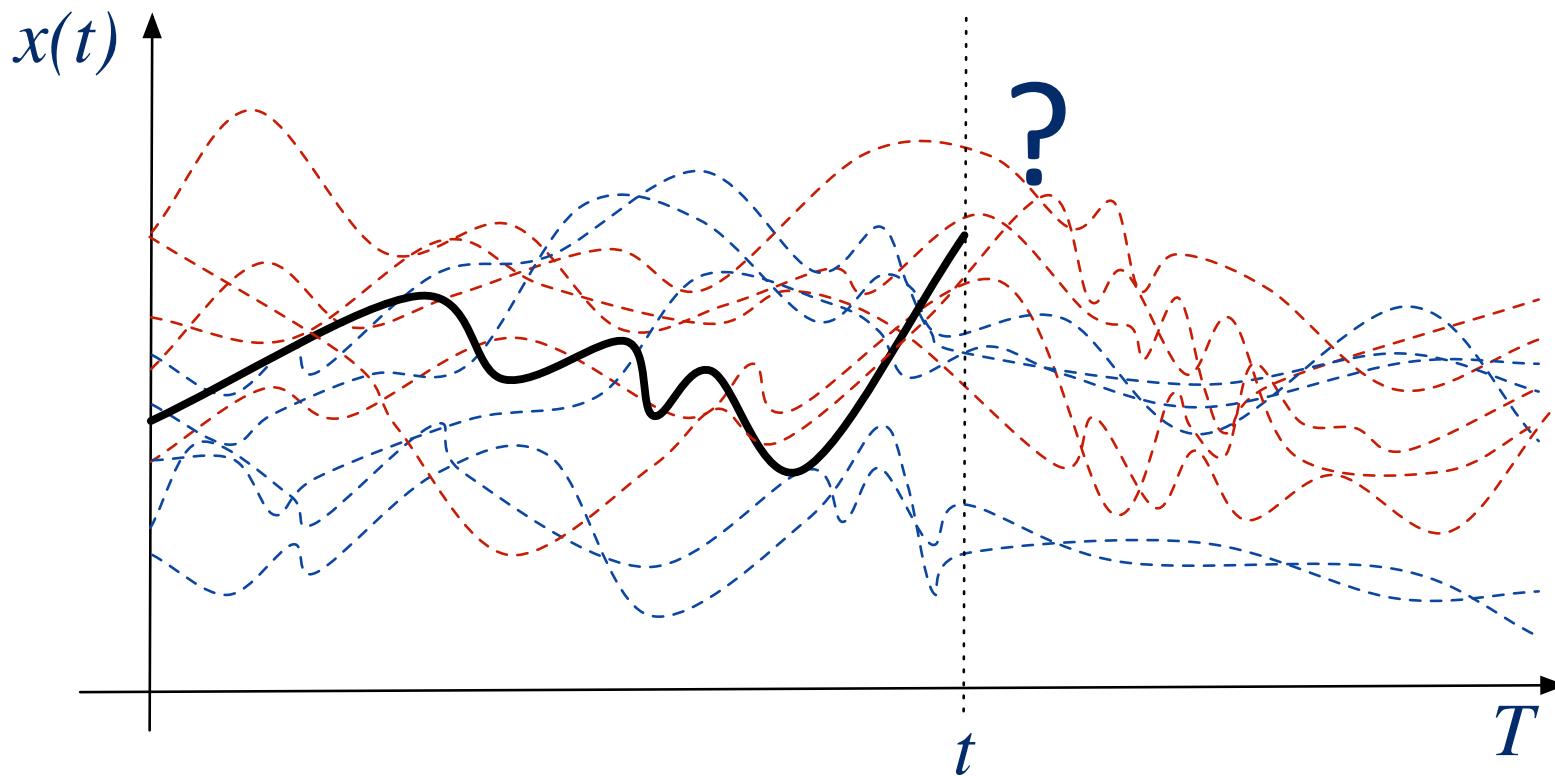
(Early) classification of time series



- Early prediction of daily **electrical consumption**: high or low

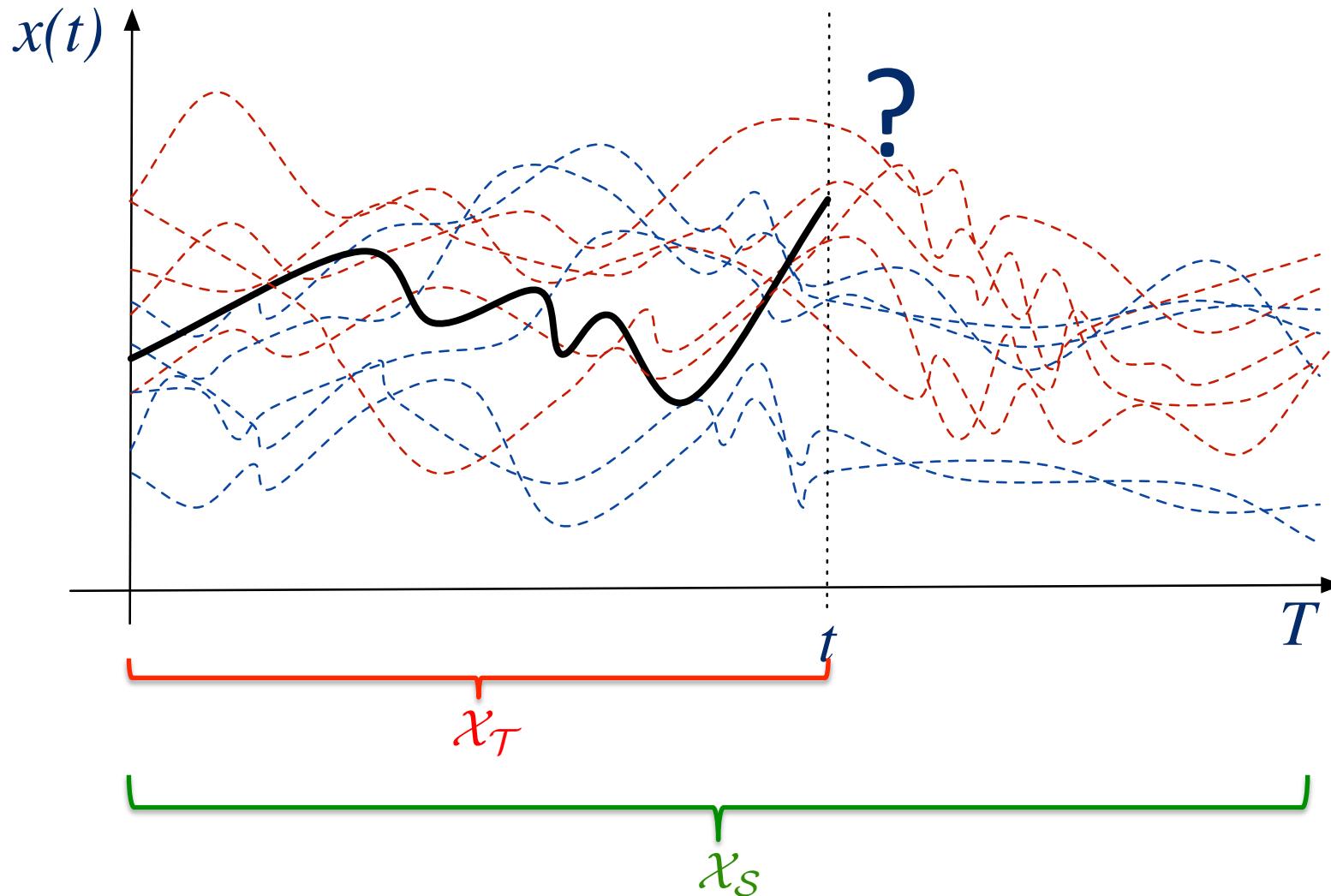
Early classification of time series

- What is the class of the new **incomplete** time series x_t ?



Early classification of time series

- What is the class of the new **incomplete** time series x_t ?



Principe

- Apprendre un **classifieur** sur les **données complètes**

$$S_{\mathcal{S}} = \{(\mathbf{x}_i^{\mathcal{S}}, y_i^{\mathcal{S}})\}_{1 \leq i \leq m} \rightarrow h_{\mathcal{S}}$$

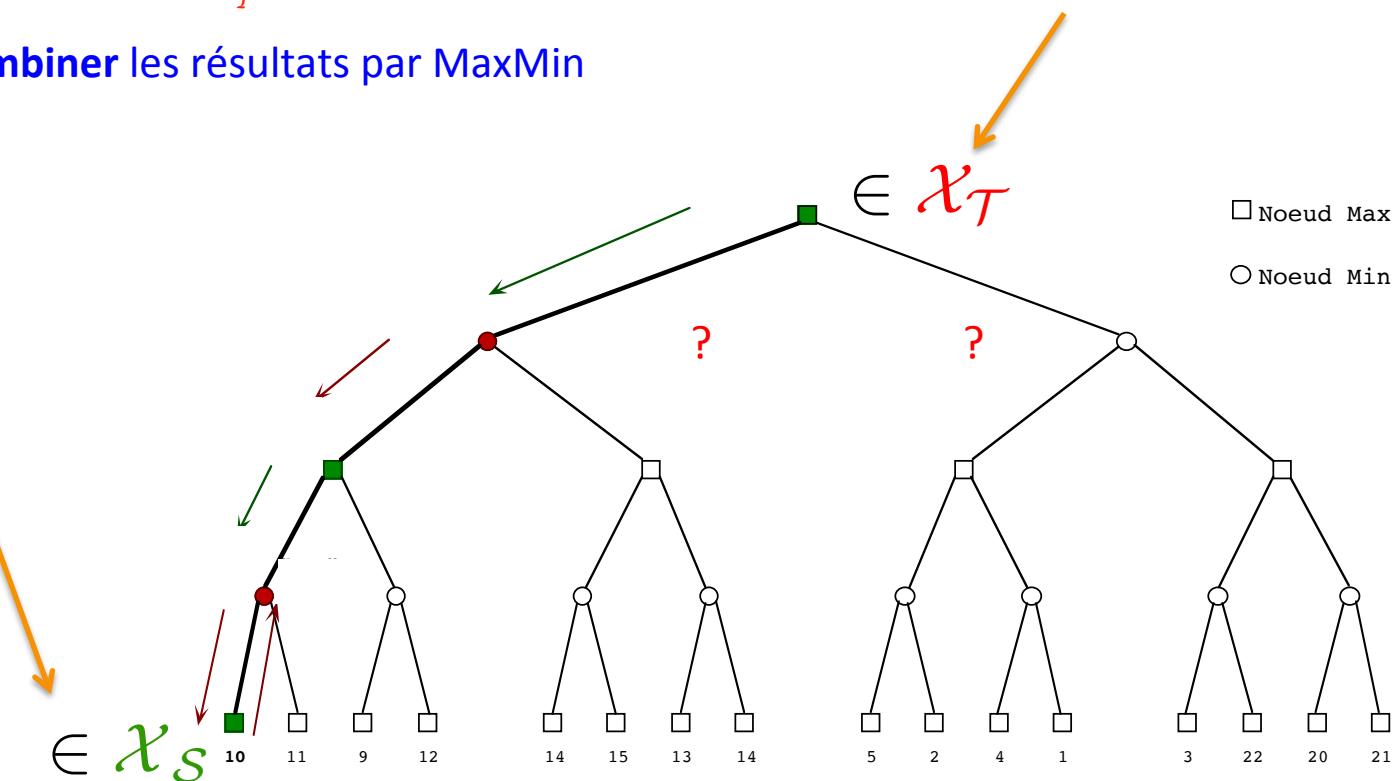
- Chercher à l'utiliser pour **classer** des **séries incomplètes**

$h_{\mathcal{T}}$ = Fonction utilisant $h_{\mathcal{S}}$

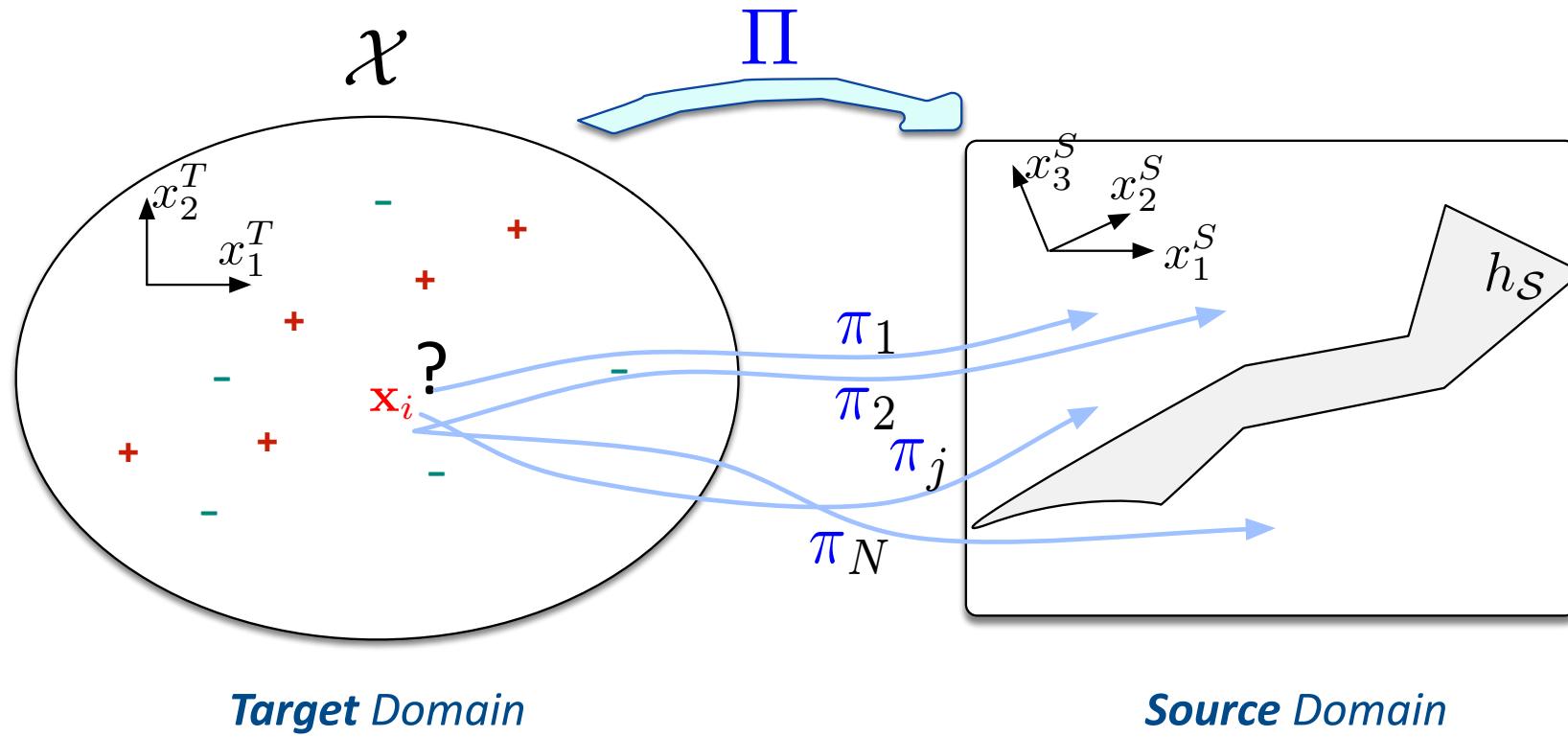
Transfert et algorithmes de jeux

Que jouer ?

- Fonction d'évaluation imparfaite dans \mathcal{X}_S
- L'utiliser dans \mathcal{X}_T
- Combiner les résultats par MaxMin



TransBoost



$$H_{\mathcal{T}}(\mathbf{x}^{\mathcal{T}}) = \text{sign} \left\{ \sum_{n=1}^N \alpha_n h_{\mathcal{S}}(\pi_n(\mathbf{x}^{\mathcal{T}})) \right\}$$

TransBoost

- Idée :
 - Apprendre des « *projections faibles* » : $\pi_i : \mathcal{X}_{\mathcal{S}} \rightarrow \mathcal{X}_{\mathcal{T}}$
 - À partir de : $S_{\mathcal{S}} = \{(\mathbf{x}_i^{\mathcal{S}}, y_i^{\mathcal{S}})\}_{1 \leq i \leq m}$
 - Par une méthode de **boosting**
 - Projection π_n telle que : $\varepsilon_n \doteq \mathbf{P}_{i \sim D_n}[h_{\mathcal{S}}(\pi_n(\mathbf{x}_i)) \neq y_i] < 0.5$
 - Re-pondération des séries temporelles d'apprentissage
 - Résultat $H_{\mathcal{T}}(\mathbf{x}^{\mathcal{T}}) = \text{sign} \left\{ \sum_{n=1}^N \alpha_n h_{\mathcal{S}}(\pi_n(\mathbf{x}^{\mathcal{T}})) \right\}$

TransBoost

Algorithm 1: Transfer learning by boosting

Input: $h_S : \mathcal{X}_S \rightarrow \mathcal{Y}_S$ the source hypothesis

$\mathcal{S}_{\mathcal{T}} = \{(\mathbf{x}_i^{\mathcal{T}}, y_i^{\mathcal{T}})\}_{1 \leq i \leq m}$: the target training set

Initialization of the distribution on the training set: $D_1(i) = 1/m$ for $i = 1, \dots, m$;

for $n = 1, \dots, N$ **do**

 Find a projection $\pi_i : \mathcal{X}_{\mathcal{T}} \rightarrow \mathcal{X}_S$ st. $h_S(\pi_i(\cdot))$ performs better than random on $D_n(\mathcal{S}_{\mathcal{T}})$;

 Let ε_n be the error rate of $h_S(\pi_i(\cdot))$ on $D_n(\mathcal{S}_{\mathcal{T}})$: $\varepsilon_n \doteq \mathbf{P}_{i \sim D_n}[h_S(\pi_n(\mathbf{x}_i)) \neq y_i]$ (with $\varepsilon_n < 0.5$) ;

 Computes $\alpha_i = \frac{1}{2} \log_2 \left(\frac{1-\varepsilon_i}{\varepsilon_i} \right)$;

 Update, for $i = 1, \dots, m$:

$$\begin{aligned} D_{n+1}(i) &= \frac{D_n(i)}{Z_n} \times \begin{cases} e^{-\alpha_n} & \text{if } h_S(\pi_n(\mathbf{x}_i^{\mathcal{T}})) = y_i^{\mathcal{T}} \\ e^{\alpha_n} & \text{if } h_S(\pi_n(\mathbf{x}_i^{\mathcal{T}})) \neq y_i^{\mathcal{T}} \end{cases} \\ &= \frac{D_n(i) \exp(-\alpha_n y_i^{(\mathcal{T})} h_S(\pi_n(\mathbf{x}_i^{(\mathcal{T})})))}{Z_n} \end{aligned}$$

 where Z_n is a normalization factor chosen so that D_{n+1} be a distribution on $\mathcal{S}_{\mathcal{T}}$;

end

Output: the final target hypothesis $H_{\mathcal{T}} : \mathcal{X}_{\mathcal{T}} \rightarrow \mathcal{Y}_{\mathcal{T}}$:

$$H_{\mathcal{T}}(\mathbf{x}^{\mathcal{T}}) = \text{sign} \left\{ \sum_{n=1}^N \alpha_n h_S(\pi_n(\mathbf{x}^{\mathcal{T}})) \right\} \quad (2)$$

Results

The diagram illustrates the experimental setup. It starts with 'Learning from target data only' (highlighted by an orange bracket), followed by 'TransBoost' (highlighted by an orange bracket), and finally 'On the source domain' (highlighted by an orange arrow) and 'Naïve transfert' (highlighted by an orange arrow). The 'Naïve transfert' row contains an orange arrow pointing to the column header H'_T (test).

slope, noise, t_T	h_T (train)	h_T (test)	H_T (train)	H_T (test)	h_S (test)	H'_T (test)
0.001, 0.001, 20	0.46 ± 0.02	0.50 ± 0.08	0.08 ± 0.03	0.08 ± 0.02	0.05	0.49 ± 0.01
0.005, 0.001, 20	0.46 ± 0.02	0.49 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01	0.45 ± 0.01
0.005, 0.002, 20	0.46 ± 0.02	0.49 ± 0.03	0.03 ± 0.02	0.04 ± 0.02	0.02	0.43 ± 0.01
0.005, 0.02, 20	0.44 ± 0.02	0.48 ± 0.03	0.09 ± 0.01	0.10 ± 0.01	0.01	0.47 ± 0.01
0.001, 0.2, 20	0.46 ± 0.02	0.50 ± 0.01	0.46 ± 0.02	0.51 ± 0.02	0.11	0.49 ± 0.01
0.01, 0.2, 20	0.42 ± 0.03	0.47 ± 0.03	0.34 ± 0.02	0.35 ± 0.02	0.02	0.35 ± 0.01
0.001, 0.001, 50	0.46 ± 0.02	0.50 ± 0.01	0.08 ± 0.03	0.08 ± 0.02	0.06	0.41 ± 0.01
0.005, 0.001, 50	0.25 ± 0.07	0.28 ± 0.09	0.01 ± 0.01	0.01 ± 0.01	0.01	0.28 ± 0.01
0.005, 0.002, 50	0.27 ± 0.07	0.30 ± 0.08	0.02 ± 0.01	0.02 ± 0.01	0.02	0.28 ± 0.01
0.005, 0.02, 50	0.26 ± 0.07	0.30 ± 0.08	0.04 ± 0.01	0.04 ± 0.01	0.01	0.31 ± 0.01
0.001, 0.2, 50	0.44 ± 0.02	0.50 ± 0.01	0.38 ± 0.03	0.44 ± 0.02	0.15	0.43 ± 0.01
0.01, 0.2, 50	0.10 ± 0.03	0.12 ± 0.04	0.10 ± 0.02	0.11 ± 0.02	0.03	0.15 ± 0.02
0.001, 0.001, 100	0.43 ± 0.03	0.47 ± 0.03	0.07 ± 0.02	0.07 ± 0.02	0.02	0.23 ± 0.01
0.005, 0.001, 100	0.06 ± 0.03	0.07 ± 0.03	0.01 ± 0.01	0.01 ± 0.01	0.01	0.07 ± 0.02
0.005, 0.002, 100	0.08 ± 0.03	0.10 ± 0.04	0.02 ± 0.01	0.02 ± 0.01	0.02	0.07 ± 0.01
0.005, 0.02, 100	0.08 ± 0.03	0.09 ± 0.03	0.02 ± 0.01	0.03 ± 0.01	0.01	0.07 ± 0.01
0.001, 0.2, 100	0.04 ± 0.03	0.46 ± 0.02	0.28 ± 0.02	0.31 ± 0.01	0.16	0.31 ± 0.01
0.01, 0.2, 100	0.03 ± 0.01	0.05 ± 0.02	0.04 ± 0.01	0.05 ± 0.01	0.02	0.05 ± 0.01

Table 1: Comparison of learning directly in the target domain (columns h_T (train) and h_T (test)), using TransBoost (columns H_T (train) and H_T (test)), learning in the source domain (column h_S (test)) and, finally, completing the time series with a SVR regression and using h_S (naïve transfer). Test errors are highlighted in the orange columns. Bold numbers indicates where TransBoost significantly dominates both learning without transfer and learning with naïve transfer.

Results

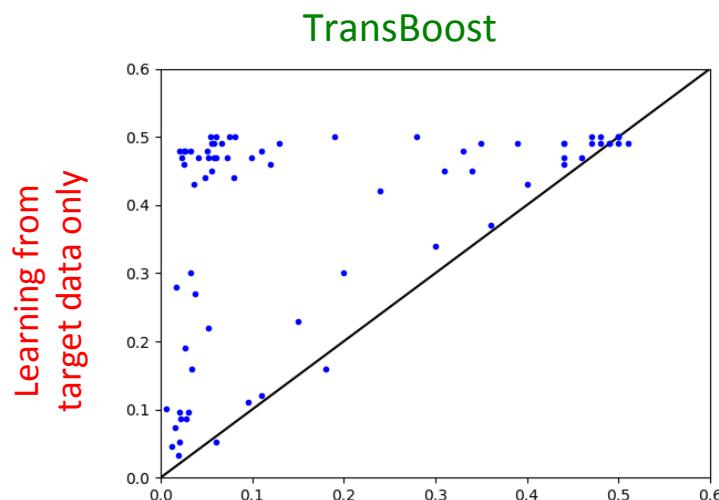


Figure 3: Comparison of error rates. y -axis: test error of the SVM classifier (without transfer). x -axis : test error of the TransBoost classifier with 10 boosting steps. The results of 75 experiments (each one repeated 100 times) are summed up in this graph.

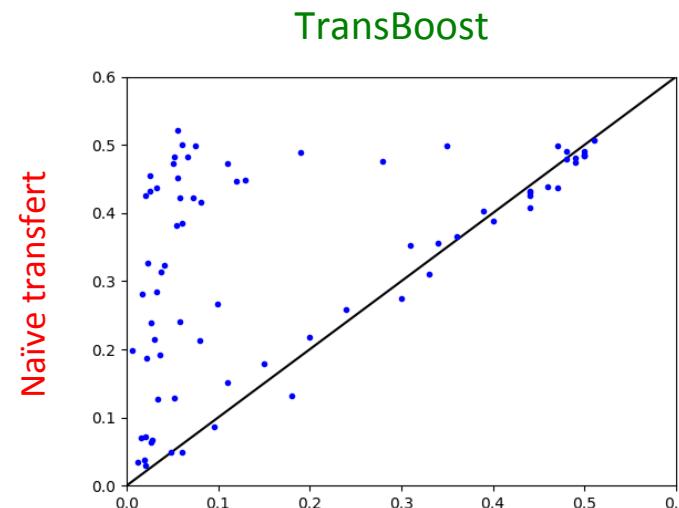
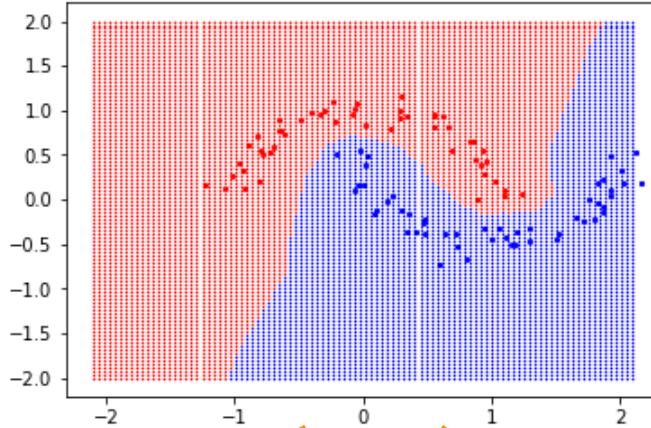


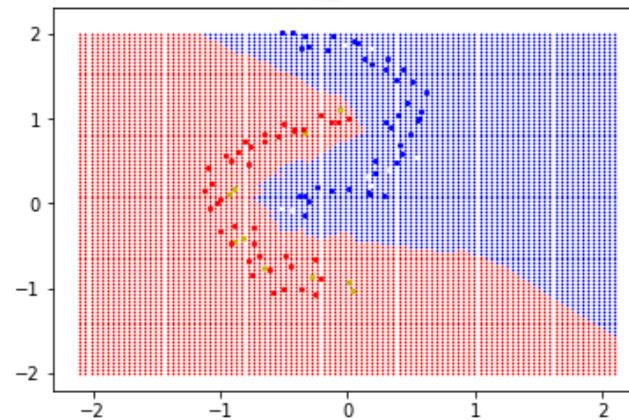
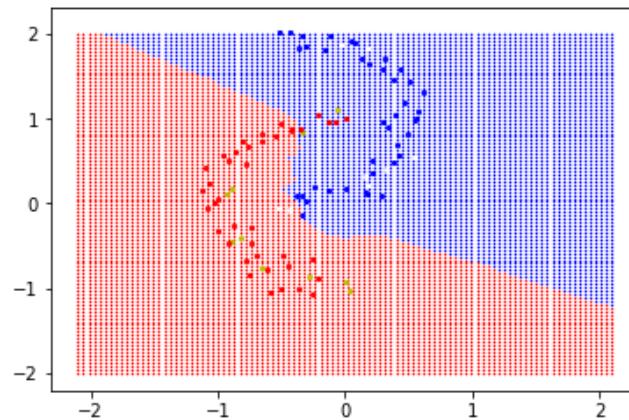
Figure 4: Comparison of error rates. y -axis: test error of the “naïve” transfer method. x -axis : test error of the TransBoost classifier with 10 boosting steps. The results of 75 experiments (each one repeated 100 times) are summed up in this graph.

Apprentissage par transfert



Apprentissage sur les données cibles
(sans transfert)

Par Transboost



Conclusion

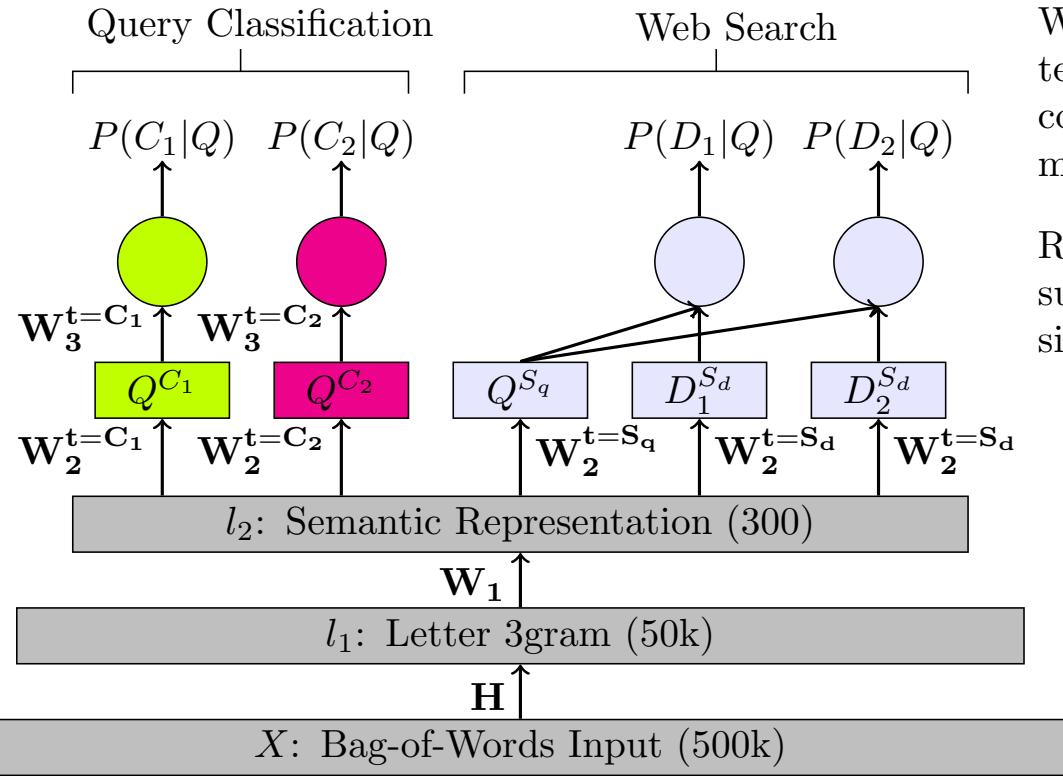
- Méthode d'ensemble et de transfert
 - Apprentissage de traducteurs faibles
 - Le problème de l'apprentissage est maintenant déplacé vers le choix d'un bon espace de projections
 - Des garanties théoriques : comme dans le boosting

Transfer with deep neural networks

Query classification posterior probability computed by sigmoid

l_3 : Task-Specific Representation (128)

Shared layers



Web search posterior probability computed by softmax

Relevance measured by cosine similarity

[X. Liu, J. Gao, X.g He, L. Deng, K. Duh and Ye-Yi Wang (2015). « *Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval* ». Proc. NAACL, May 2015]

Bilan sur l'apprentissage supervisé : pourquoi ça marche

- On peut évaluer la performance des experts

Et donc ...

- Les **sélectionner**
- Les **construire** pour qu'ils soient complémentaires (boosting)
- Les **combiner** (e.g. vote pondéré)

Plan

1. Contexte et motivations
2. Méthodes collaboratives en IA
3. Méthodes collaboratives en Apprentissage Automatique
4. Quid du clustering ?
5. Bilan

Méthodes collaboratives en Clustering

Grands scénarios et motivations

1. Clustering « coopératif » : Recherche d'une **solution consensus**

- Experts « **faibles** » : diminution de la variance
- Expertises différentes et **complémentaires** (e.g. blackboard)

Grands scénarios et motivations

2. Clustering « collaboratif » :

Calcul de **solutions locales** avec échanges entre les « experts »

- **Informations** utiles disponibles chez les autres collaborateurs
- **Perturbation** de l'**exploration** de l'espace des solutions : échapper aux minima locaux
- **Modification** du **biais** local incertain

Les questions

1. Clustering « **coopératif** » : Recherche d'une **solution consensus**
 1. Comment **générer** des solutions « faibles » ?
 2. Quelle information **transmettre** ?
 3. Comment **combiner** les informations ?

Les questions

1. Clustering « coopératif » : Recherche d'une **solution consensus**
 1. Comment **générer** des solutions « faibles » ?
 2. Quelle information **transmettre** ?
 3. Comment **combiner** les informations ?
2. Clustering « collaboratif » :
Calcul de **solutions locales** avec échanges entre les « experts »
 1. Les collaborateurs existent : comment en **sélectionner / les pondérer** ?
 2. Quelle information **transmettre** ?
 3. Comment **combiner** les informations ?

Les questions

1. Clustering « coopératif » : Recherche d'une **solution consensus**

1. Comment **générer** des solutions « faibles » ?
2. Quelle information **transmettre** ? 
3. Comment **combiner** les informations ?

2. Clustering « collaboratif » :

Calcul de **solutions locales** avec échanges entre les « experts »

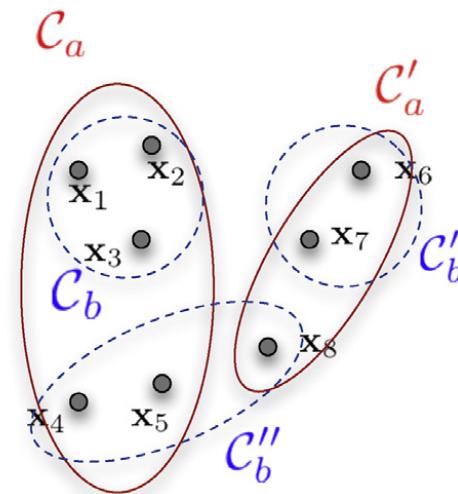
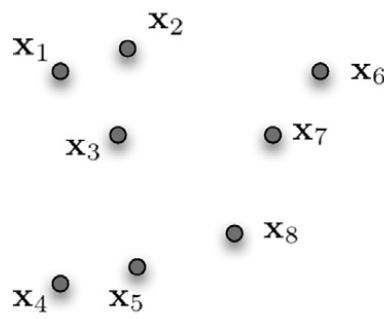
1. Les collaborateurs existent : comment en **sélectionner / les pondérer** ?
2. Quelle information **transmettre** ? 
3. Comment **combiner** les informations ?

Quels échanges d'information ?

1. Mêmes **données** / même **espace**
 - Attribution des **données** aux clusters
2. Mêmes **données** / **espaces** différents
 - Attribution des **données** aux clusters
3. **Données** différentes / même **espace**
 - Comparaison des **solutions**
 - Centroïdes ; nombre de clusters ; ...
4. **Données** différentes / **espaces** différents
 - Comparaison des **solutions**
 - Nombre de clusters ; ...

Mêmes données / même espace ou espaces différents

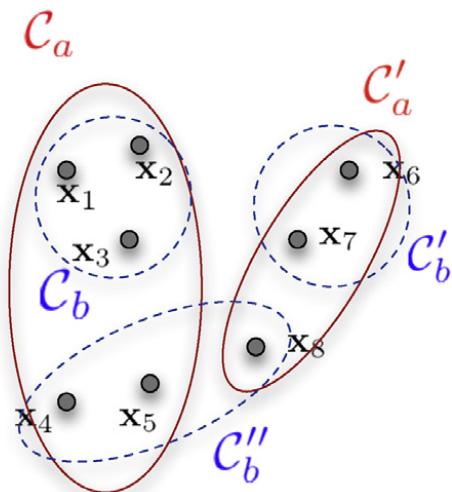
- Illustration



Algorithm	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_7	\mathbf{x}_8
A	\mathcal{C}_a	\mathcal{C}_a	\mathcal{C}_a	\mathcal{C}_a	\mathcal{C}_a	\mathcal{C}'_a	\mathcal{C}'_a	\mathcal{C}'_a
B	\mathcal{C}_b	\mathcal{C}_b	\mathcal{C}_b	\mathcal{C}'_b	\mathcal{C}'_b	\mathcal{C}''_b	\mathcal{C}''_b	\mathcal{C}''_b

Mêmes données / même espace

- Recherche de correspondance entre les clusters



	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_7	\mathbf{x}_8
\mathbf{x}_1	-	1	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
\mathbf{x}_2	1	-	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
\mathbf{x}_3	1	1	-	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
\mathbf{x}_4	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	-	1	0	0	$\frac{1}{2}$
\mathbf{x}_5	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	-	0	0	$\frac{1}{2}$
\mathbf{x}_6	0	0	0	0	0	-	1	$\frac{1}{2}$
\mathbf{x}_7	0	0	0	0	0	1	-	$\frac{1}{2}$
\mathbf{x}_8	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	-

Matrice de consensus ou d'association

Données différentes

- Communication sur les **solutions** trouvées
 - Centroïdes
 - Gaussiennes
 - Nombre de clusters

Grands scénarios et approches

- Recherche d'une **solution consensus**
 - Solution « **barycentrique** »
 - Définition d'un critère à optimiser
 - Définition d'une **mesure de similarité**
 - Procédure
 - **Calcul direct**
 - **Itérative** (type k-Means) avec espace de description approprié
 - **Partition** d'une matrice de similarité
 - Et clustering hiérarchique
 - Par **vote**

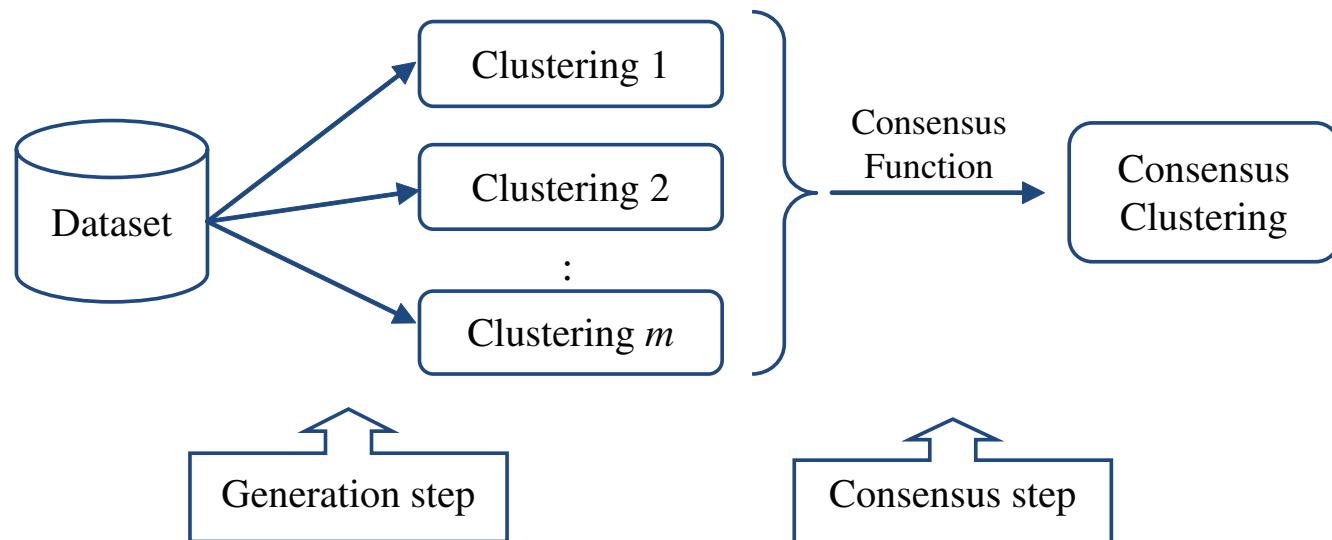
Grands scénarios et approches

- Calcul de solutions **locales** avec échanges entre les « experts »
 - Éventuellement **sélection / pondération** des collaborateurs
 - **Combinaison locale des avis**
 - Passage par une **solution barycentrique**
 - **Critère d'optimisation local**
- 
- Définition d'une
mesure de similarité

Recherche de consensus

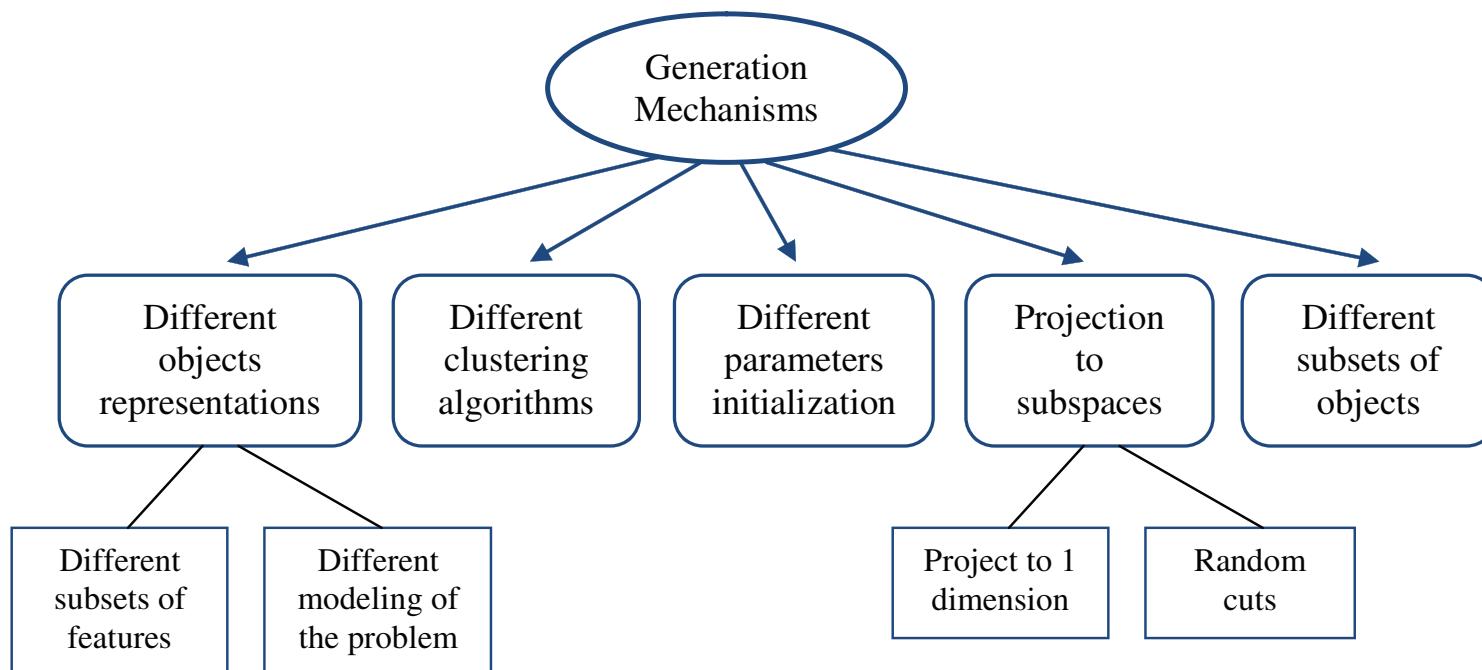
Recherche de consensus

- Principe



Recherche de consensus

- **Génération de solutions**



[Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms.

International Journal of Pattern Recognition and Artificial Intelligence, 25(03), 337-372.]

Combinaison des solutions

Soit $\mathcal{S} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ l'ensemble des N objets à catégoriser

Soit $\mathbb{C} = \{\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(J)}\}$ un ensemble de J résultats de clustering de \mathcal{S}

- On définit : $\Phi : \mathbb{C} \rightarrow \mathcal{C}^*$

tel que

$$\mathcal{C}^* = \underset{\mathcal{C}}{\text{ArgMax}} \left\{ \sum_{j=1}^J \text{sim}(\mathcal{C}, \mathcal{C}^{(j)}) \right\}$$

Sorte de moyenne ou de barycentre des clusterings



Il faut définir la notion de « similarité »

Mesures de similarité

- *Information mutuelle*

- Quantifie l'information statistique entre deux distributions

$$\mathcal{S} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$$

Une partition de \mathcal{S} : $U = \{U_1, U_2, \dots, U_K\}$ et une autre partition $V = \{V_1, V_2, \dots, V_{K'}\}$.

La probabilité qu'un objet tiré aléatoirement dans \mathcal{S} tombe dans le cluster U_i

est : $P(i) = \frac{|U_i|}{N}$.

De même : $P'(j) = |V_j|/N$

$$MI(U, V) = \sum_{i=1}^K \sum_{j=1}^{K'} P(i, j) \log \frac{P(i, j)}{P(i)P'(j)}$$

$$P(i, j) = \frac{|U_i \cap V_j|}{N}$$

Méthodes

$$\mathcal{C}^* = \operatorname{ArgMax}_{\mathcal{C}} \left\{ \sum_{j=1}^J sim(\mathcal{C}, \mathcal{C}^{(j)}) \right\}$$

- Quand la mesure de similarité est symétrique et $J > 2$,
le problème est **NP-difficile** (variantes encore à l'étude)

[M. Krivanek and J. Moravek, Hard problems in hierarchical-tree clustering,
Acta Inform. 3 (1998) 311323.]



Recours à des méthodes heuristiques

Méthodes de recherche de solution consensus

- Optimisation heuristique
 - Recuit simulé ; Algorithmes Génétiques
 - ...
- Autres approches sans optimisation directe
 - Basée sur la matrice d'association
 - Plus un exemple appartient aux mêmes clusters dans les solutions locales, plus il a de chance d'appartenir à ce cluster dans la solution consensus
 - Et méthode de clustering hiérarchique
 - Basées sur la définition d'un nouvel espace de définition induit par les clusterings solutions
 - Passage dans un espace de dimension N (nombre d'objets)
 - Et algorithme de type k-Means dans cet espace

	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4	\mathbf{x}_5	\mathbf{x}_6	\mathbf{x}_7	\mathbf{x}_8
\mathbf{x}_1	-	1	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
\mathbf{x}_2	1	-	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
\mathbf{x}_3	1	1	-	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
\mathbf{x}_4	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	-	1	0	0	$\frac{1}{2}$
\mathbf{x}_5	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	-	0	0	$\frac{1}{2}$
\mathbf{x}_6	0	0	0	0	0	-	1	$\frac{1}{2}$
\mathbf{x}_7	0	0	0	0	0	1	-	$\frac{1}{2}$
\mathbf{x}_8	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	-

$$\mathbf{Y} = \{y_1, y_2, \dots, y_N\}$$

$$y_i = \langle \pi_1(\mathbf{x}_i), \dots, \pi_J(\mathbf{x}_i) \rangle$$

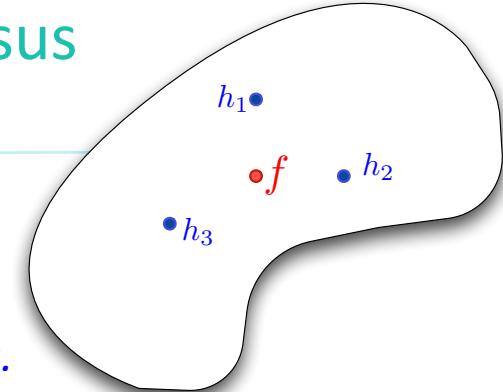
[Nguyen, N., & Caruana, R. (2007). **Consensus clusterings**. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (pp. 607-612). IEEE.]

Recherche d'une solution consensus

Par combinaison d'experts faibles

The strongest argument in favour of cluster ensembles is as follows.

1. *It is known that the current off-the-shelf clustering methods may suggest very different structures in the same data. This is the result of the **different clustering criteria** being optimized.*
2. *There is no layman guide to choosing a clustering method for a given data set and so an inexperienced user runs the risk of picking an inappropriate clustering method. There is **no ground truth** against which the result can be matched, therefore there is no critique to the user's choice.*
3. *Cluster ensembles provide **a more universal solution** in that various structures and shapes of clusters present in data may be discovered by the same ensemble method, and the solution is less dependent upon the chosen ensemble type.*



[Hadjitodorov et al. (2006) « Moderate diversity for better cluster ensembles ».
Information Fusion, 7 (2006), pp.264-275]

Bilan

1. Pas de preuve

sur l'amélioration du clustering consensus / aux clusterings sources

[Topchy, A., Jain, A. K., & Punch, W. (2003, November). Combining multiple weak clusterings. In *3rd IEEE Int. Conf. on Data Mining* (pp. 331-338). IEEE.]

Bilan

1. Pas de preuve

sur l'amélioration du clustering consensus / aux clusterings sources

2. Avantage (?)

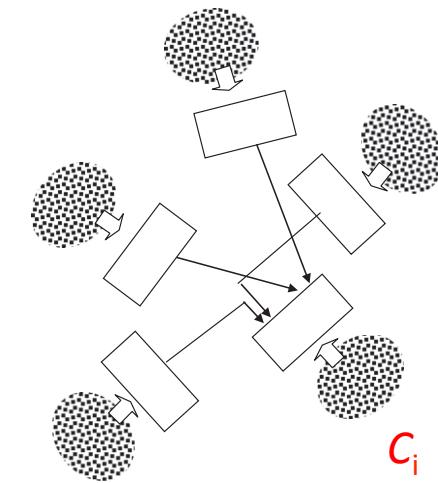
Éviter un choix d'algorithme (de biais) à l'utilisateur

[Topchy, A., Jain, A. K., & Punch, W. (2003, November). Combining multiple weak clusterings. In *3rd IEEE Int. Conf. on Data Mining* (pp. 331-338). IEEE.]

Recherche de solutions locales

Clustering collaboratif : Calcul de solutions locales

L'agent i peut **sélectionner** ses interlocuteurs
et les **pondérer**



$$C_i^*(\mathcal{S}_i) = \operatorname{Argmin}_{\mathcal{C}(\mathcal{S}_i)} \left\{ \underbrace{f(\mathcal{C}(\mathcal{S}_i))}_{\text{local criterion}} + \underbrace{g \left[\text{agreement} \left(\mathcal{C}(\mathcal{S}_i), \{\mathcal{C}_j \mid 1 \leq j \leq J, j \neq i\} \right) \right]}_{\text{collaborative criterion}} \right\}$$

Clustering collaboratif

$$\mathcal{C}_i^*(\mathcal{S}_i) = \underset{\mathcal{C}(\mathcal{S}_i)}{\operatorname{Argmin}} \left\{ \underbrace{f(\mathcal{C}(\mathcal{S}_i))}_{\text{local criterion}} + \underbrace{g \left[\text{agreement}(\mathcal{C}(\mathcal{S}_i), \{\mathcal{C}_j \mid 1 \leq j \leq J, j \neq i\}) \right]}_{\text{collaborative criterion}} \right\}$$

- Mesure de l'accord entre solutions
 - Complexité de Kolmogorov [Murena et al. (2018)]
 - Basée sur l'entropie [Sublime et al. (2017)] : méthodes probabilistes
- Approche
 - En deux étapes (répétées)
 1. Calcul des **solutions locales**
 2. Calcul des solutions **prenant en compte** les autres solutions
 - De manière itérative : synchrone ou asynchrone

Clustering collaboratif : similarité

- Critère d'optimisation
 - Passage par une solution barycentrique
 - Fondée sur la complexité de Kolmogorov [Murena et al. (2018)]

$$\mathcal{C}^* = \operatorname{Argmin}_{\mathcal{C}, \mathcal{C}(\mathcal{S}_i) (i=1, \dots, J)} \left\{ \underbrace{\sum_{j=1}^J \left[K(\mathcal{C}(\mathcal{S}_j)) + K(\mathcal{S}_j | \mathcal{C}(\mathcal{S}_j)) \right]}_{\text{local criterion}} + \underbrace{K(\mathcal{C}) + \sum_{j=1}^J K(\mathcal{C}(\mathcal{S}_j) | \mathcal{C})}_{\text{collaborative criterion}} \right\}$$

Clustering collaboratif : sélectionner ses collaborateurs

- En **apprentissage supervisé**, on sait comment sélectionner des collaborateurs car on sait évaluer leur qualité
- En **apprentissage non supervisé** ...
 - Notion de “bon clustering”
 - **solution stable** 
 - Contre de petites **perturbations** des **données**
 - Contre de petites **perturbations** de **biais**

$$\mathcal{C}_i^*(\mathcal{S}_i) = \operatorname{ArgMax}_{\mathcal{C}(\mathcal{S}_i)(i=1,\dots,J)} \left\{ \underbrace{\sum_{j=1}^J \textcolor{red}{f}(\mathcal{C}(\mathcal{S}_j))}_{\text{local criterion}} + \underbrace{\sum_{j=1, j \neq i}^J \tau_{i,j} \Delta(\mathcal{C}(\mathcal{S}_i), \mathcal{C}_j)}_{\text{collaborative criterion}} \right\}$$

$$\forall i : \quad \sum_{i=1, i \neq j}^J (\tau_{i,j})^p = 1, \quad p \in \mathbb{N}^*$$

Sélectionner et pondérer ses collaborateurs

- Les solutions
 - Qui favorisent les **collaborateurs qui sont d'accord**
 - Seraient de bonnes solutions (clusterings)
- Mais
 - « Chambres d'écho »
 - Ne tient pas compte de l'**indépendance des algorithmes**

[Cornuéjols, A., & Martin, C. (2014). **Une méthode d'ensemble en apprentissage non supervisé quand on ne connaît rien sur la performance des experts.** *Revue des Nouvelles Technologies de l'Information (RNTI)*, 2016, vol. RNTI-A-8, pp. 33-50.]

Plan

1. Contexte et motivations
2. Méthodes collaboratives en IA
3. Méthodes collaboratives en Apprentissage Automatique
4. Quid du clustering ?
5. Bilan

- Mêmes **algorithmes** sur des **données** différentes
 - Intéressant si **les données partagent quelque chose** sur leurs distributions
 - **Mesure de similarité** entre les données ???
 - Forcément en fonction d'un modèle
- Mêmes **données** => il faut des **algorithmes** différents
 - **Sélection**
 - **Construction** (à la Boosting)
- **Données** différentes et **algorithmes** différents
 - ???

Quel lien avec les méthodes collaboratives en IA et en AA ?

- **Méthodes d'ensemble** : combinaison de solutions faibles
 - Sous-ensemble de données
 - Initialisations différentes
 - Biais différents
- Mais **pas** d'évaluation « objective » de la **performance**

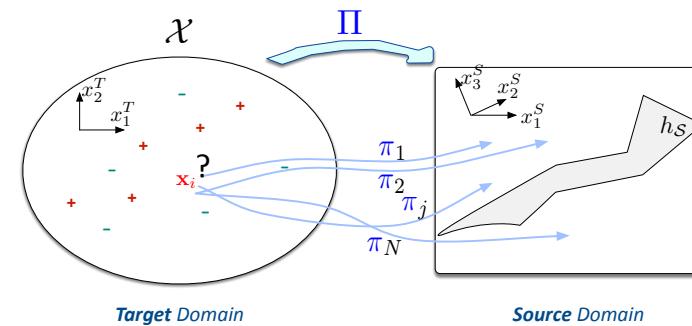
} Idem

Il manque une théorie
analogue aux théories en apprentissage supervisé

Quel lien avec les méthodes collaboratives en IA et en AA ?

- Calcul de solutions locales

- Sélection des « collaborateurs » pertinents
- Apprendre à traduire ?
(comme Transboost)



Quel lien avec les méthodes collaboratives en IA et en AA ?

- Calcul de solutions locales
 - Sélection des « collaborateurs » pertinents
 - Apprendre à traduire ?
(comme Transboost)
- Nombreux systèmes heuristiques
 - Mais il faut encore d'autres idées
 - Et une théorie

Le clustering collaboratif est un **problème d'avenir**

Références

- Cornuéjols, A., Wemmert, C., Gançarski, P., & Bennani, Y. (2018). Collaborative clustering: Why, when, what and how. *Information Fusion*, 39, 81-95.
- Depaire, B., Falcón, R., Vanhoof, K., & Wets, G. (2011). PSO driven collaborative clustering: A clustering algorithm for ubiquitous environments. *Intelligent Data Analysis*, 15(1), 49-68.
- Domeniconi, C., & Al-Razgan, M. (2009). Weighted cluster ensembles: Methods and analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(4), 17.
- Murena, P. A., Sublime, J., Matei, B., & Cornuéjols, A. (2018, July). An Information Theory based Approach to Multisource Clustering. In *IJCAI* (pp. 2581-2587).
- Nguyen, N., & Caruana, R. (2007, October). Consensus clusterings. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (pp. 607-612). IEEE.
- Pedrycz, W. (2002). Collaborative fuzzy clustering. *Pattern Recognition Letters*, 23(14), 1675-1686.
- Qiao Y., Li S., Denœux T. (2019) « Collaborative Evidential Clustering ». In: Kearfott R., Batyrshin I., Reformat M., Ceberio M., Kreinovich V. (eds) *Fuzzy Techniques: Theory and Applications. IFSA/NAFIPS 2019 2019. Advances in Intelligent Systems and Computing*, vol 1000. Springer, Cham
- Sublime, J., Matei, B., Cabanes, G., Grozavu, N., Bennani, Y., & Cornuéjols, A. (2017). Entropy based probabilistic collaborative clustering. *Pattern Recognition*, 72, 144-157.
- Topchy, A., Jain, A. K., & Punch, W. (2003, November). Combining multiple weak clusterings. In *Third IEEE International Conference on Data Mining* (pp. 331-338). IEEE.
- Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03), 337-372.