Ma recette santé !
Étude expérimentale du transfert analogique

Balkis Bouthaina Dirahoui, Sarah Djouder, Tinhinane Chafai
Encadré par : Marie-Jeanne Lesot

2 juin 2022

Table des matières

1. Transfert Analogique

2. Implémentation
   - Architecture globale
   - Modélisation
   - CoAT optimisé

3. Expérimentations
   - Validation CCBI
   - Comparaison du temps d'exécution : CCBI vs Knn,SVM
   - Temps d'excution : CoAT vs CCBI
   - Performance CCBI : PREC, CONF, MAE

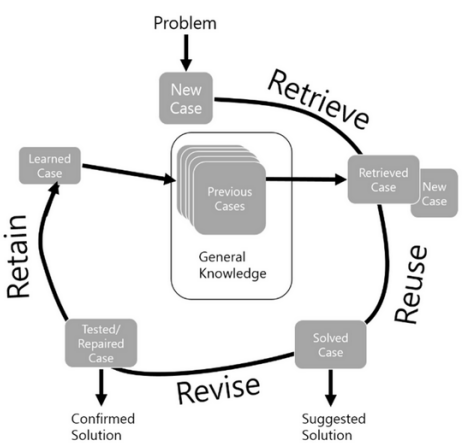
4. Application à des données réelles
   - Traitements
   - Classfication
   - Régression

Introduction : Transfert Analogique

Définition :

- Méthode d'apprentissage supervisé qui permet de résoudre des problèmes de régression ou de classification
- Semblable au modèle des K plus proches voisins...

- Stockage des données d'apprentissage.
- Utilise ses données pour une nouvelle prédiction.
- Utilisation de ces données pour une nouvelle prédiction
- Travail **principalement** fait à **l'inférence**.



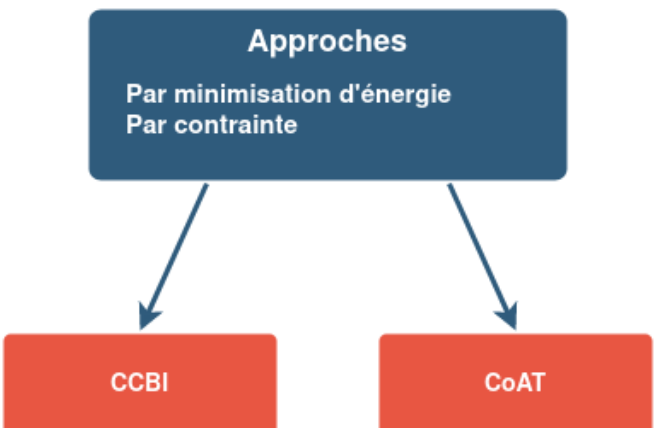
1

1. By Sangsun Jung al. Sustainability 2020

Introduction : Transfert Analogique

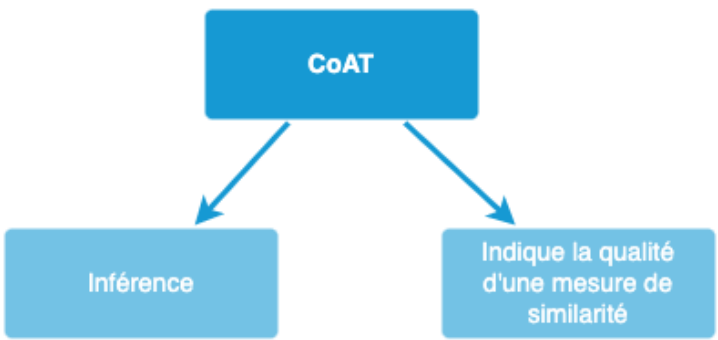
Hypothèse

- Si deux situations sont similaires dans un aspect alors elles sont aussi similaires dans d'autres aspects.



Une approche ordinale : CoAT [2]

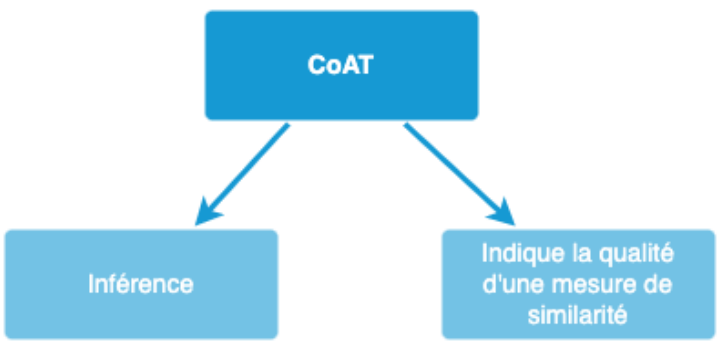
- Traduit le principe d'analogie sous forme **ordinale**.
- Introduit le concept d'**inversion**.
- Introduit le concept de **complexité**.



2. Fadi Badra. 2020. A Dataset Complexity Measure for Analogical Transfer. In Proceedings of the 29th International Joint Conference on Artificial Intelligence IJCAI 2020, 1601–1607.

Une approche ordinale : CoAT

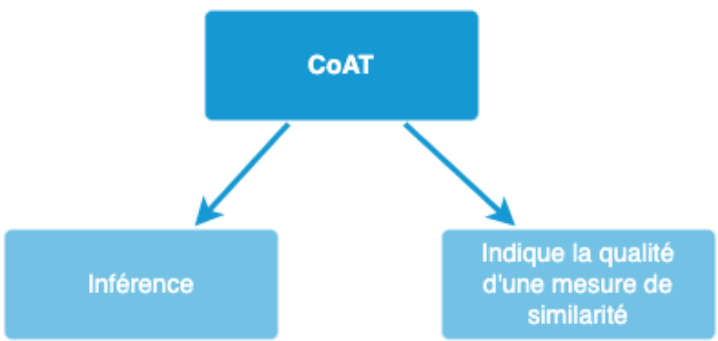
- **Notations**
  - **triplet** (s_0, s_i, s_j)
  - **mesure de similarité** σ_s
  - **mesure de similarité** σ_r



Une approche ordinale : CoAT

- **Notion d'inversion**

$$Inv(s_0) = \{s_i s_j | \sigma_s(s_0, s_i) \geq \sigma_s(s_0, s_j) \; et \; \sigma_r(s_0, s_i) < \sigma_r(s_0, s_j)\}$$



Une approche ordinale : CoAT

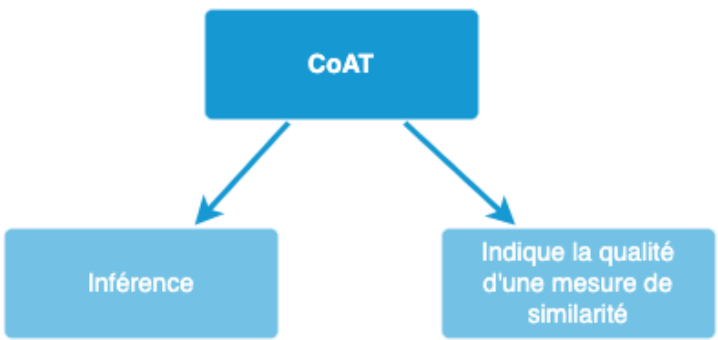
- **Cardinal**

$$\gamma(s_0) = |inv(s_0)|$$

- **Complexité** Γ

$$\Gamma(\sigma_s, \sigma_r, CB) = \sum_{s_0 \in CB} \gamma(s_0)$$

Complexité temporelle en $O(n^3)$!!

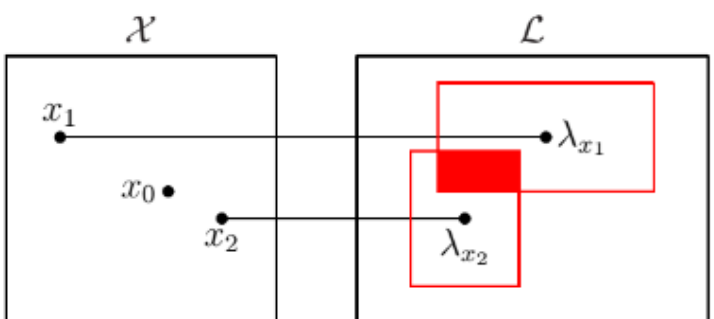


Approche par contrainte : CCBI [3]

- **Différence :** Pas des triplets $(s_0, s_i, s_j) \rightarrow$ **couples** (x_i, λ_{xi})
- Contrainte de similarité :

$$\forall x, y \in X : simX(x, y) \leq simL(\lambda_x, \lambda_y)$$

- Un cas (x_1, λ_{x1}) soit $simX(x1, x0) = \alpha 1$ **solution** $\lambda_{x0} \in \alpha 1$-voisinage de λ_{x1}



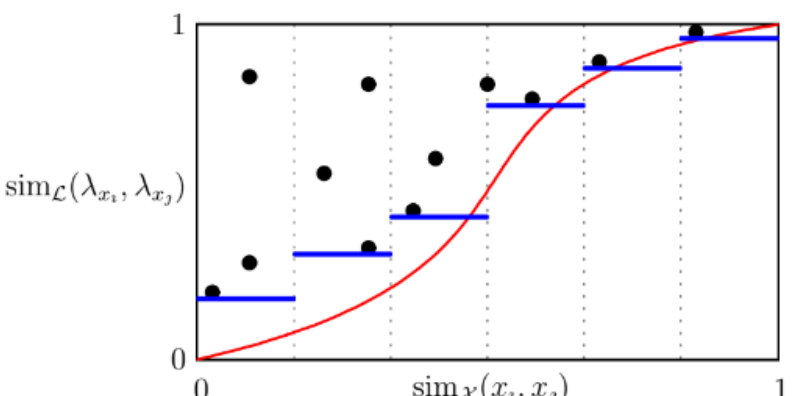
3. Eyke Hullermeier. 2007. Credible Case-Based Inference Using Similarity Profiles. IEEE Trans. Knowl. Data Eng. 19, 6 (2007), 847–858.

Approche par contrainte : CCBI

- Une contrainte de similarité :

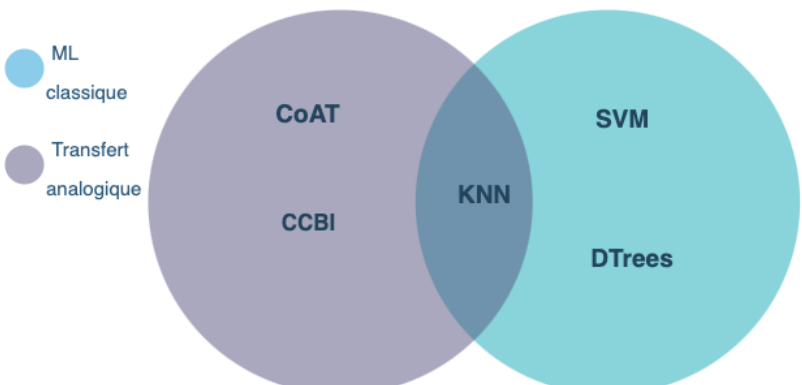
$$\forall x, y \in X : simX(x, y) \leq simL(\lambda_x, \lambda_y)$$

- Apprentissage d'une fonction $h : x \mapsto \sum_{k=1}^{n} \beta_k . \mathbb{1}_{(A_k)}(x)$



Implémentation : Architecture Globale

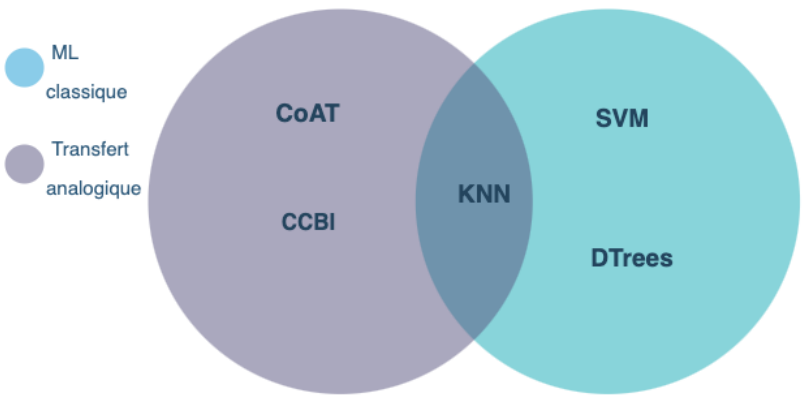
- **Méthodes disponibles** : CoAT,CCBI, KNN,SVM,DTrees...
- **Paramètres de chaque modèle :**
  - k KNN
  - β_k CCBI...
- **Mesures disponibles :**
  - SimX
  - SimL
  - σ_s
  - Distance euclidienne
  - **autre**



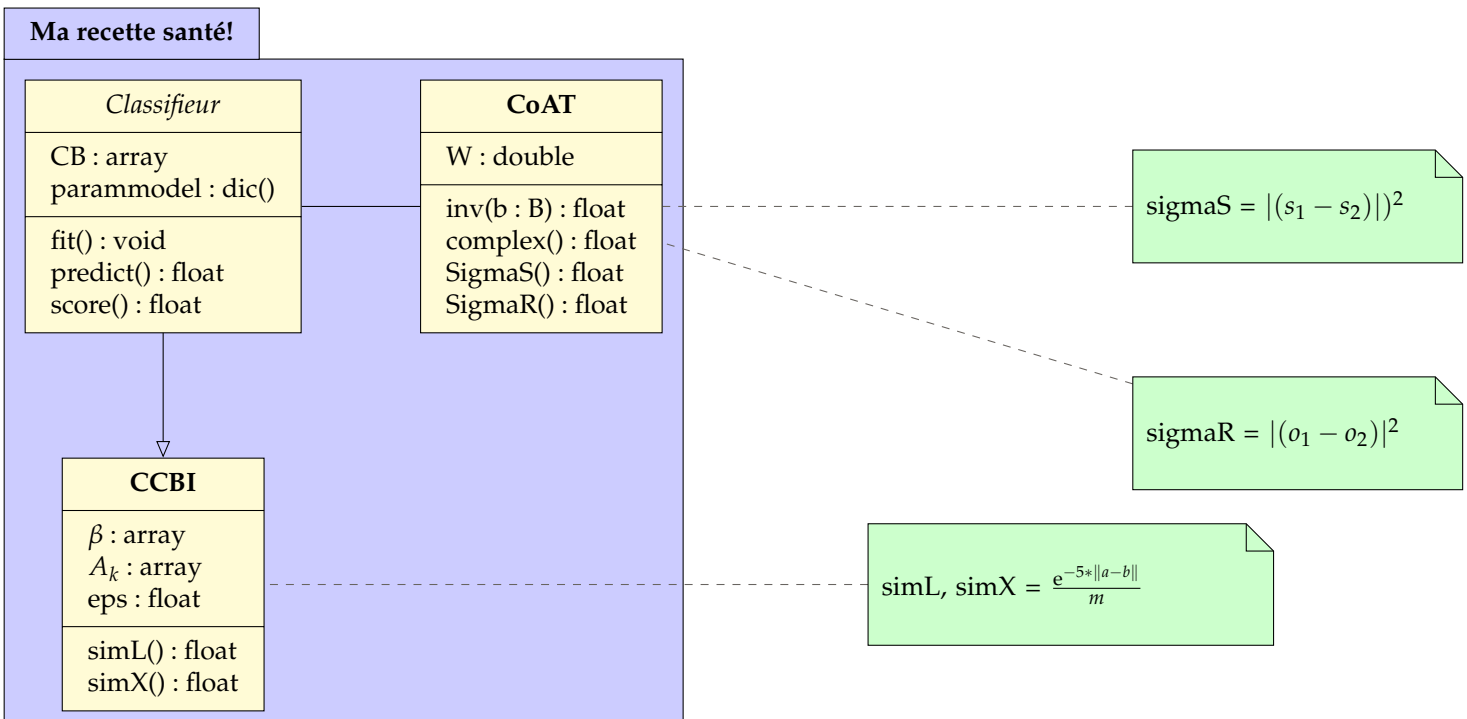
Implémentation : Architecture Globale

Pour la plate-forme de comparaison expérimentale :

- **Évaluation des performances**
  - Temps d'exécution
  - Validation croisée
  - Split test-train
- **Paramètres d'apprentissage :**
  - Nombre d'exemples en test
  - k en validation croisée.



Implémentation : Modélisation



Implémentation : CoAT optimisé

Rappelez vous...

- **Minimiser** la complexité :

$$\Gamma(\sigma_s, \sigma_r, CB) = \sum_{s_0 \in CB} \gamma(s_0)$$

Implémentation : CoAT optimisé

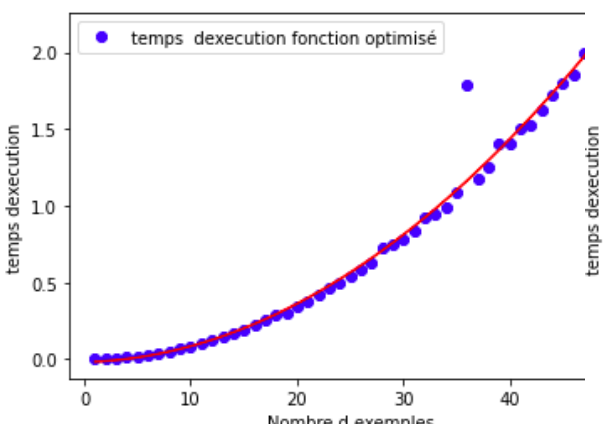
Et si ? :

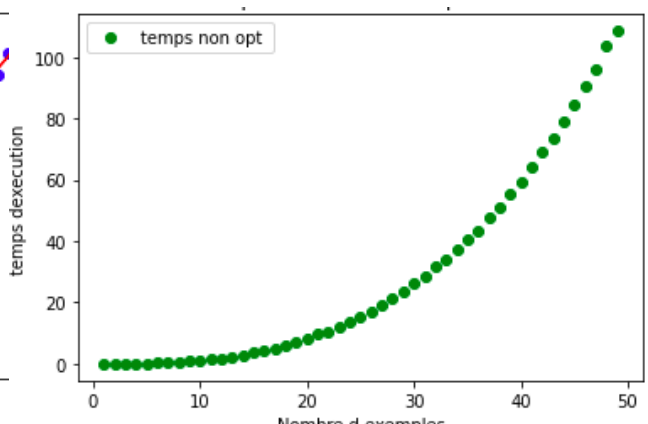
- En inférence : calculer pour tout le dataset ?
- Calculer **seulement** $|inv(s_0)|$.
- Optimiser le temps de calcul..

$$\Rightarrow \text{Peut être réduit de } O(n^3) \text{ à } O(n^2)$$

Expérimentation : Performances de la version optimisée

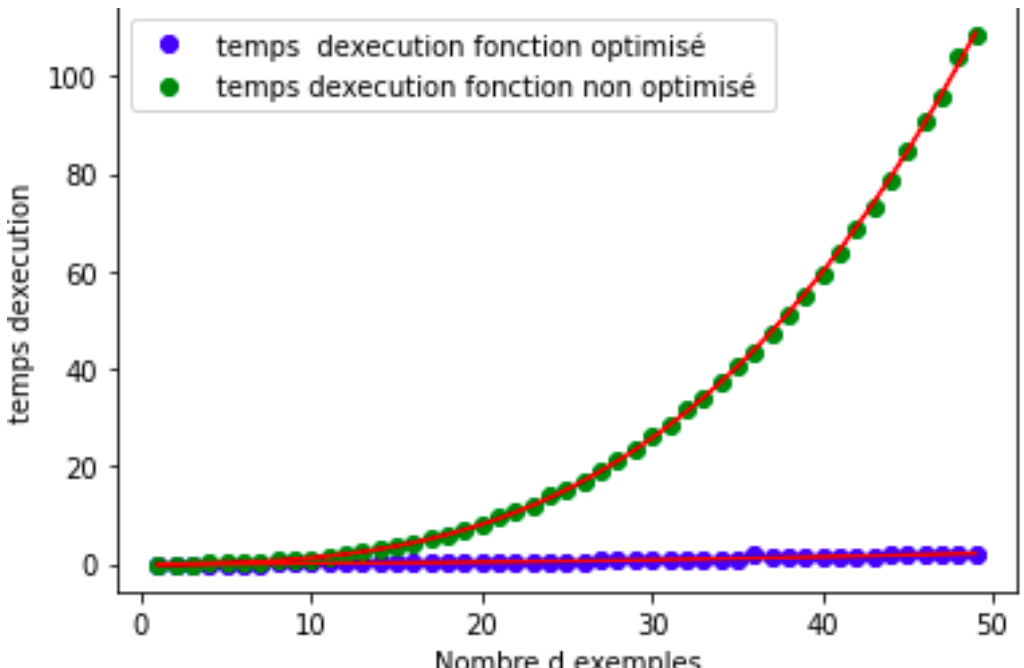
- Non-optimisée (vert) en $O(n^3)$
- Optimisée (bleu) en $O(n^2)$





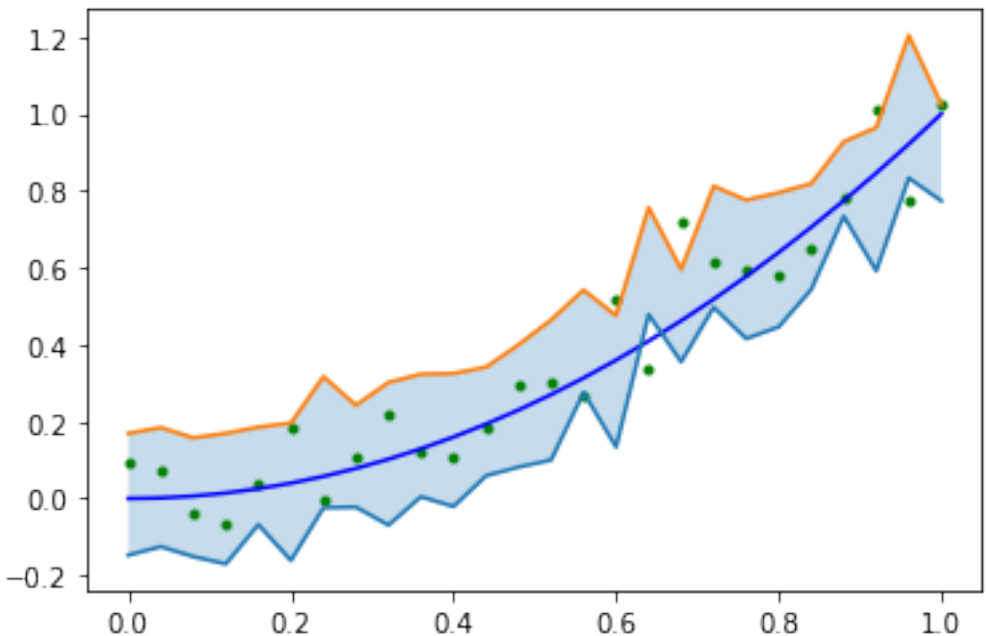
Expérimentation : Performances de la version optimisée

- Non-optimisée (vert) en $O(n^3)$
- Optimisée (bleu) en $O(n^2)$



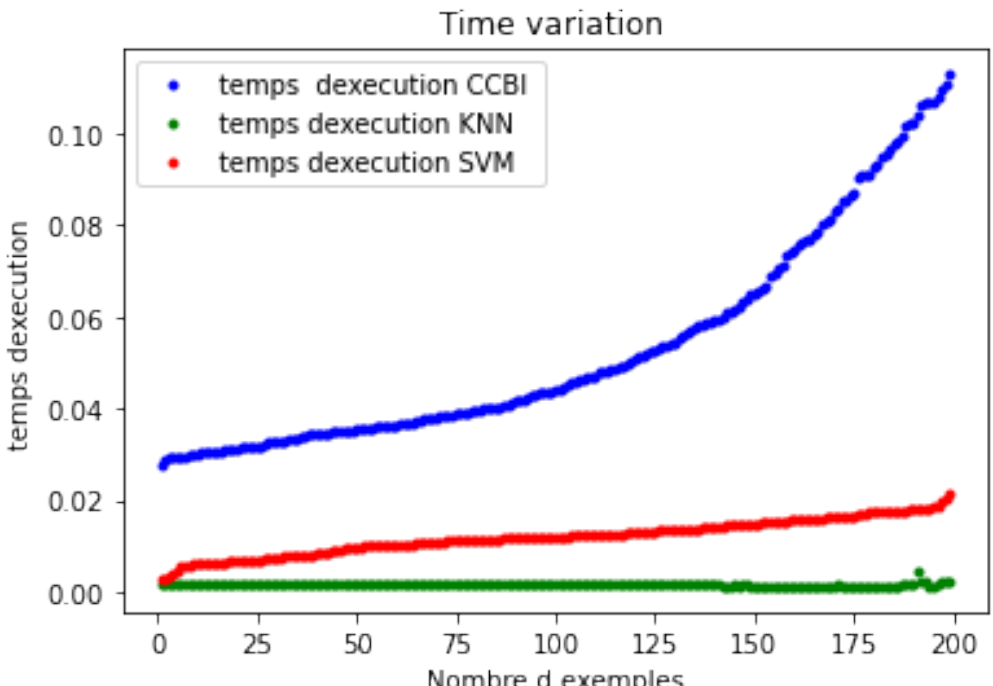
Validation des résultats CCBI

- Ligne bleue : $x \to x^2$
- Points verts : Exemple d'apprentissage bruités $x \to x^2$
- Intervalle prédit (région entre les lignes brisées)



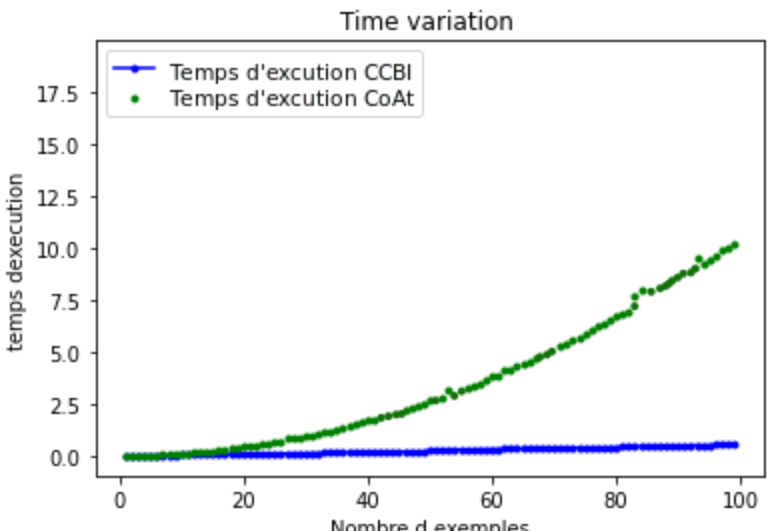
Temps d'exécution : CCBI vs Knn, SVM

- Ligne bleue : $x \rightarrow x^2$
- Temps d'exécution par rapport au nombre d'exemples



Temps d'exécution : CoAT vs CCBI

- Problème de régression : $x \rightarrow x^2$
- **Différence** : double boucle 'inv()' utilisé dans **CoAT**
- Ligne verte : CoAT
- Ligne bleue : CCBI



Performance CoAT/CCBI vs SVM, Knn, Dtrees

Modèle	SVM	Knn	Dtrees	CoAT
Score (Accuracy)	0.98	0.98	0.90	0.98

Performances des modèles sur les données IRIS

Modèle	SVM	Knn	Dtrees	CCBI
Score	0.98	0.99	0.98	0.94

Performances des modèles sur les données x^2

- Ici la mesure utilisée pour CCBI est la confiance (CONFIDENCE)
- CONF : Taux de bonne classification $y \in intervalle$.

Performance : CONF, PREC, MAE

- Validation de l'algorithme CCBI sur des données synthétiques
- **Mpg (resp. Cpu)** : Données sur les véhicules (resp. Données sur les ordinateurs) de la base UCI [4]
- **CONF** : Taux de bonne classification $y \in intervalle$.
- **PREC** : Taille de l'intervalle prédit.
- **MAE** : Mean Absolute Error.

Mesure	CONF	PREC	MAE
mpg	0.94	0.06	0.01
cpu	0.98	0.35	0.15

4. https ://archive.ics.uci.edu/ml/datasets.php

Application à des données réelles

- Données recettes de cuisine issues du site BBC Good Food
- Sous format Json

{"page": {"article": {"author": "Jane Hornby", "description": "Give traditional roast chicken a Chinese-style twist with this zesty recipe", "id": "101919", "tags": []}, "recipe": {"collections": ["Broccoli", "Winter roasts", "Chinese chicken", "Chicken leg"], "cooking_time": 2700, "prep_time": 600, "serves": 4, "keywords": ["Chinese", "Lunch", "Family", "Roast", "One pot", "Supper", "Super", "Dinner", "Diner", "BBC Good Food magazine July", "Broccoli", "Chicken leg", "Chicken legs", "Clear honey", "Garlic clove", "Garlic cloves", "Ginger", "Lemon", "Lemons", "Light soy sauce", "Sesame oil", "Sesame seed", "Sesame seeds"], "ratings": 87, "nutrition_info": ["Added sugar 6g", "Carbohydrate 7g", "Kcal 340 calories", "Protein 28g", "Salt 2.35g", "Saturated fat 7g", "Fat 23g"], "ingredients": ["chicken leg", "lemon", "broccoli", "garlic clove", "ginger", "light soy sauce", "clear honey", "sesame seed", "sesame oil"], "courses": ["Main course", "Supper", "Lunch"], "cusine": "Chinese", "diet_types": [], "skill_level": "Easy", "post_dates": "1309474800"}, "channel": "Recipe", "title": "Lemon, broccoli & sesame roast chicken"}}

Extrait d'une recette du Dataset Json

- Environ 3000 recettes avec 30 attributs différents pour chacune

Traitements

- **Implémentation d'un parser** :
  1. Expressions régulières
  2. Extractions des valeurs nutritionnelles .
- **Ingrédients : sac de mots**
  1. Tokenization
  2. Stemming
  3. TF-IDF
- **Binarisation des valeurs catégorielles**

Jeu de Données : Caractéristiques du jeu de données

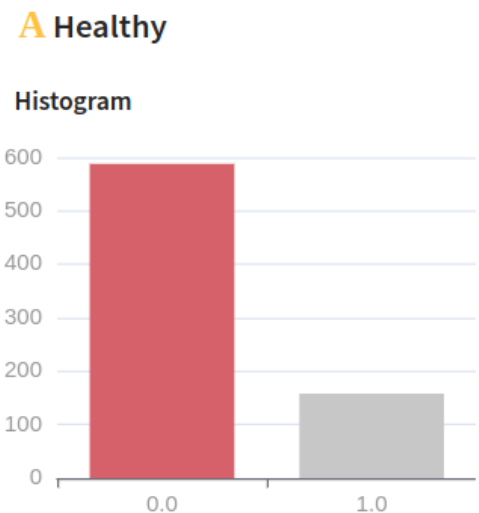
Caractéristiques du jeu de donnée

1. Plus 1000 recettes
2. 49 attributs différents (ingrédients,sucre,fibres,...)
3. 5 tâches de classification
4. 2 tâches de régression

Classification : La recette est-elle saine ou non ?

Protocole Expérimental

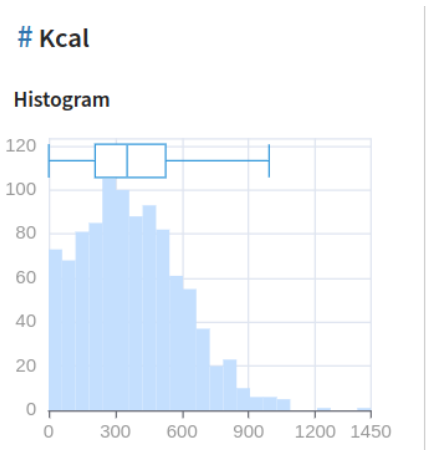
- **Équilibrer les classes**
  - UnderSampling
- **Séparer les données**
  - train , test



Modèle	SVM	Knn	Dtrees	CoAT
Score (Accuracy)	0.72	0.82	0.78	0.80

Performances des modèles en classification sur les données BBC Good Food

Régression : quel est l'apport calorique de la recette ?



Distribution des calories

Modèle	LR	CCBI	CoAT
R2-Score	0.82	0.80	0.91

Performances des modèles en régression sur les données BBC Good Food

Conclusion

1. **Réalisations** :
   - Implémentation d'approches du transfert analogique (CCBI + CoAT)
   - Optimisation théorique et pratique de CoAT
   - Développement d'une plateforme de comparasion expérimentale
   - Construction d'un jeu de données réelles
   - Application à des données réelles

Conclusion

1. **Réalisations** :
   - Implémentation d'approches du transfert analogique (CCBI + CoAT)
   - Optimisation théorique et pratique de CoAT
   - Développement d'une plateforme de comparasion expérimentale
   - Construction d'un jeu de données réelles
2. **Bilan :**
   - L'idée du transfert analogique est séduisante mais :
   - Cout de Calcul prohibitf
   - Pertinent pour des petits jeux de données.