# Les forêts d'arbres extrêmement aléatoires : utilisation dans un cadre non supervisé

Kevin Dalleau*, Miguel Couceiro*
Malika Smail-Tabbone*

*Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

**Résumé.** Dans ce travail, nous présentons une nouvelle méthode permettant le calcul de similarités entre objets basée sur les forêts d'arbres extrêmement aléatoires. L'idée principale de notre méthode est de séparer les données de manière itérative jusqu'à ce qu'une condition d'arrêt soit respectée, et de calculer une similarité basée sur la co-occurrence des instances dans les feuilles de chaque arbre obtenu. Nous évaluons la méthode sur un ensemble de jeux de données synthétiques et réels. Cette évaluation est basée sur la comparaison des similarités moyennes entre instances ayant la même étiquette aux similarités moyennes entre instances d'étiquette différente. Ces mesures sont comparables aux notions de similarités intracluster et intercluster, mais ont pour intérêt d'être agnostiques aux choix d'une méthode de clustering en particulier. L'étude empirique montre que la méthode permet effectivement de distinguer les individus n'appartenant pas aux même clusters. Les forêts d'arbres extrêmement aléatoires non supervisées ont des propriétés intéressantes, telles que : (i) l'invariance aux transformations monotones de variables, (ii) la robustesse aux variables corrélées, et (iii), la robustesse au bruit. Enfin, nous présentons les résulats obtenus par l'applicaton d'un algorithme de clustering hiérarchique agglomératif, en utilisant les matrices de similarité obtenues par notre méthode. Les résultats obtenus sur des jeux de données homogènes et hétérogènes sont prometteurs.

## 1 Introduction

De nombreux algorithmes d'apprentissage non supervisé se basent sur une mesure de similarité ou distance entre instances. Bien qu'il existe un grand nombre de métriques décrites dans la littérature, en pratique, l'ensemble de métriques disponibles est grandement réduit par les caractéristiques des données et de l'algorithme choisi. Le choix d'une distance peut impacter fortement la qualité d'un clustering.

Shi et Horvath proposent dans (Shi et Horvath (2006)) la méthode des *Unsupervised Random Forest* (URF), dérivant des random forests (RF, Breiman (2001)). Leur méthode permet de calculer des distances entre instances non étiquetées en utilisant les forêts d'abres aléatoires. Une fois la forêt construite, les données d'entraînement sont passées dans chacun des arbres. Chaque feuille contenant un nombre limité d'instances, et toutes les instances terminant dans les mêmes feuilles pouvant être considérées comme similaires, il est possible de définir une

mesure de similarité : si deux objets $i$ et $j$ sont dans la même feuille, la similarité globale entre ces deux objets est incrémentée. Cette mesure est par la suite normalisée via une division par le nombre d'arbres dans la forêt. L'utilisation des forêts d'arbres aléatoires est rendue possible dans le cadre non supervisé grâce à la génération d'instances synthétiques, permettant la classification binaire entre ces dernières et les instances observées, non étiquetées. Deux méthode de génération d'instance sont présentées dans leur travaux, *addCl1* et *addCl2*.

L'utilisation d'arbres de décision en tant que méthode pour obtenir une similarité permet de limiter les prétraitements nécessaires, notamment dans le cas de jeux de données hétérogènes.

Malgré ses avantages, la méthode présente deux principaux inconvénients. Premièrement, la phase de génération d'instances n'est pas efficace sur la plan computationnel. Dans la mesure où les arbres obtenus dépendent des instances générées, il est nécessaire de constuire plusieurs forêts avec différentes instances générées et d'effectuer une moyenne de leurs résultats. Deuxièmement, les instances synthétiques peuvent biaiser le modèle construit vers la discrimination d'objets sur certains attibuts spécifiques.

En parallèle, P. Geurts, D. Ernst et L. Wehenkel présentent dans Geurts et al. (2006) un nouveau type de méthode d'ensemble à base d'arbres de décision nommé *Extremely Randomized Trees*, ou *ExtraTrees* (ET). Cet algorithme est similiaire à l'algorithme des RF sur de nombreux points. Dans les forêts d'arbres aléatoires, l'échantillonnage concerne les instances et les attributs. Dans les ET, les seuils de coupure à chaque noeud sont eux aussi obtenus de manière partiellement ou complètement aléatoire. À chaque noeud, $K$ attributs sont aléatoirement sélectionnées et une coupure aléatoire est réalisée. La meilleure coupure est gardée.

Il est intéressant de développer deux paramètres de l'algorithme des ET : $K$, et $n_{min}$, la taille minimale d'un noeud pour qu'il puisse subir une coupure. Le paramètre $K$, ayant des valeurs dans l'ensemble $\{1, ..., n_{features}\}$, influence le caractère aléatoire des arbres. En effet, pour des valeurs faibles, la dépendance des arbres vis-à-vis des étiquettes est diminuée. Dans le cas extrême où $K = 1$ (*i.e.* uniquement un attribut est sélectionné pour la réalisation de la coupure), la dépendance entre les arbres et les étiquettes est éliminée.

Nous proposons d'étendre l'approche de Shi et Horvath (2006) et d'utiliser les ET avec une nouvelle approche où la génération d'instances n'est plus nécessaire, que nous nommons *unsupervised extremely randomized trees* (UET). Nous étendons cette méthode afin de pouvoir l'utiliser sur des données hétérogènes.

Ce document est un résumé du travail présenté à PAKDD 2018 (Dalleau et al. (2018)).

## 2 Unsupervised Extremely Randomized Trees

Les méthodes de génération d'instances synthétiques ne sont pas efficace computationnellement. Au lieu de générer de nouvelles instances, une autre approche possible est d'assigner aléatoirement de nouvelles étiquettes. Cette méthode que nous proposons et nommons *addCl3*, peut-être implémentée de la manière suivante. Soit $n_{obs}$ le nombre d'instances dans le jeu de données. Une liste contenant $\lfloor \frac{n_{obs}}{2} \rfloor$ fois l'étiquette 0 et $n_{obs} - \lfloor \frac{n_{obs}}{2} \rfloor$ l'étiquette 1.

Pour chaque instance, une étiquette est échantillonée sans remplacement dans cette liste. Avec *addCl3*, deux instances similaires peuvent être étiquetées différement et finir dans deux feuilles différentes. Cependant, en fixant $K = 1$, la construction des arbres ne dépend plus de l'étiquette. L'algorithme des ET semble donc être un algorithme intéressant à utiliser avec *addCl3*. L'algorithme 1 présente les UET.

### 2.1 Description de l'algorithme

**Algorithme 1 :** Unsupervised Extremely Randomized Trees

**Data :** Observations $O$
**Result :** Matrice de similarité S
$D \longleftarrow addCl3(O)$;
$T \longleftarrow Build\_an\_extra\_tree\_ensemble(D, K)$ // *Ici* $K = 1$;
$S = 0_{n_{obs}, n_{obs}}$ // *Initialisation d'une matrice nulle* $n_{obs}$ ;
**for** $d_i \in D$ **do**
  **for** $d_j \in D$ **do**
    $S_{i,j}$ = nombre de fois où les échantillons $d_i$ et $d_j$ sont dans la même feuille ;
  **end**
**end**
$S_{i,j} = \frac{S_{i,j}}{M}$;

La procédure Build_an_extra_tree_ensemble(D,K) est celle présentée dans Geurts et al. (2006).

L'utilisation de méthode à base d'abres permet l'application à des données hétérogènes, les attributs continus, catégoriels et ordinaux étant traités. Il est à noter que la procédure de génération d'étiquettes, *addCl3*, n'est pas nécessaire. En effet, ces étiquettes ne portant pas d'information sur les instances et $K = 1$, chaque coupure est réalisée sans considération de l'étiquette.

### 2.2 Optimisation des paramètres

Deux paramètres sont importants : le nombre d'arbres $n_{trees}$, et $n_{min}$, vu précédemment. Nous avons évalué l'influence de ces paramètres sur 3 jeux de données, *Iris*, *Wine* and *Wisconsin* [1]. Notre évaluation, basée sur l'évolution de l'Adjusted Rand Index (ARI), montre que (i) la différence d'ARI n'est pas statistiquement significative pour $n_{trees} > 50$ et (ii) la valeur optimale de $n_{min}$ semble être autour de $30\%$ de la taille du jeu de données. La significativité des différences a été évaluée par le test de Kruskall-Wallis.

## 3 Évaluation empirique

### 3.1 Évaluation de certaines caractéristiques des UET

Nous avons comparé la différence entre la similarité moyenne entre les instances appartenant au même groupe et la similarité moyenne entre instances n'appartenant pas au meme groupe, que l'on dénote $\Delta$. La procédure est répétée 20 fois. Nous calculons $\bar{\Delta}$, la moyenne des $\Delta$, ainsi que $\sigma$, l'écart type. Cette approche permet une comparaison agnostique à une méthode de clustering donnée.

1. Tous ces jeux de données sont disponibles sur l'UCI Machine learning repository, `https://archive.ics.uci.edu/ml/datasets.html`

| Dataset | $\bar{\Delta}$ | $\sigma$ |
|---|---|---|
| *NoC4* | 0.00042 | 0.00003 |
| *NoC50* | 0.00007 | 0.00003 |
| *C4* | 0.68417 | 0.00341 |

TAB. 1 – $\bar{\Delta}$ *dans différents jeux de données avec et sans clusters.*

**Capacité à discriminer des clusters** Nous avons généré trois jeux de données pour cette expérience : deux jeux de données sans structure de cluster, *NoC4* et *NoC50*, ainsi qu'un jeu de données avec une structure de clustern *C4*. Les résultats, présentés Table 3.1, montre que notre méthode semble être capable de donner des $\bar{\Delta}$ significativement plus grands lorsqu'une structure de cluster existe, tout en donnant des $\bar{\Delta} \approx 0$ lorsqu'il n'y en a pas.

**Invariance aux transformations monotones d'attributs** L'une des propriétés intéressantes des arbres de décision est leur invariance à ce type de transformations. En effet, comme précisé dans Friedman (2006), cette propriété confère une résistance aux *outliers*, ainsi qu'à tout changement d'échelle entre variables. Une évaluation empirique sur deux jeux de données synthétiques semble confirmer cette propriété.

**Robustesse aux variables corrélées** Nos expérimentations semblent indiquer que les UET sont robustes à la présence de variables corrélées. Cette propriété est intéressante dans la mesure où elle permet de limiter les étapes de prétraitement dans le cas où des variables fortement corrélées sont présentes dans les données.

## 3.2 Performance des UET sur des données numériques, catégorielles et hétérogènes

Nous avons appliqué le même protocole à des jeux de données synthétiques et réels de 3 types différents : continu, catégoriel, et hétérogènes. Pour les 6 jeux de données d'évaluation, contenant entre $4$ et $10$ variables et $2$ classes, $0.30 < \bar{\Delta} < 0.49$, et $0.001 < \sigma < 0.008$.

## 3.3 Évaluation comparative avec des résultats de la littérature

Dans les sous-sections précédentes, nous avons comparé les différences de similarité intra- et inter- cluster. Il est aussi intérssant de comparer les résultats obtenus avec de véritables clusterings présentés dans la littérature. Le protocole suivant est adopté dans cette partie. Pour chaque jeu de données, UET est appliqué 10 fois, et la moyenne des matrices de similarité est calculée. Cette matrice est par la suite transformée en matrice de distance par l'équation $DIS_{ij} = \sqrt{1 - SIM_{ij}}$ et un clustering hiérarchique agglomératif est réalise. La qualité du clustering est évalué par la *Normalized Mutual Information* (NMI). Cette procédure est répétée 20 fois, et la moyenne et l'écart type de la NMI sont calculés. Les résltats obtenus sont présentés Table 3.3.

| Jeu de donnée | UET - NMI | Littérature - NMI |
|---|---|---|
| Wisconsin | **78.33** $\pm$ 3.25 | 73.61 $\pm$ 0.00 |
| Lung | **29.98** $\pm$ 6.17 | 22.51 $\pm$ 5.58 |
| Breast tissue | **74.48** $\pm$ 2.92 | 51.18 $\pm$ 1.38 |
| Isolet | 61.22 $\pm$ 1.47 | **69.83** $\pm$ 1.74 |
| Parkinson | **25.50** $\pm$ 6.14 | **23.35** $\pm$ 0.19 |
| Ionosphere | **13.47** $\pm$ 1.11 | **12.62** $\pm$ 2.37 |
| Segmentation | **69.62** $\pm$ 2.14 | 60.73 $\pm$ 1.71 |

TAB. 2 – *Évaluation Comparative avec les résultats de Elghazel et Aussem (2010). Les meilleurs résultats sont indiqués en gras. Les deux valeurs sont indiquées en gras dans les cas d'égalité.*

Nous avons par ailleurs comparé les résultats obtenus avec la méthode proposée et les URF. Pour ce faire, nous avons utilisé l'implémentation proposée par les auteurs et comparés les ARI obtenues en utilisant l'algorihtme de partitionnement autour des médoïdes. 2000 arbres et 100 forêts sont utilisés pour les URF, avec une valeur de $m_{try} = \lfloor \sqrt{n_{features}} \rfloor$. Les matrices de similarités sont obtenues par moyennage de 20 matrices de similarité. Ces expériences ont été réalisées sur une machine équipée d'un processeur Intel i7-6600U. Tout en obtenant des clusterings compétitifs avec la littérature, un gain concernant le temps de calcul est parfois observé, parfois de plusieurs ordres de grandeur.

### 3.4 Comparaison avec la distance euclidienne

Enfin, nous avons comparé les clusterings obtenus en utilisant les matrices obtenues par notre méthode avec des clusterings obtenus en utilisant la distance euclidienne. Pour les jeux de données categoriels, la distance euclidienne est calculée après application de *One-Hot encoding* afin de transformer les variables. Une comparaison de la NMI montre que les UET sont performants face à la distance euclidienne, donnant des NMI meilleures ou comparables dans tous les cas testés.

## 4 Conclusion et perspectives

Dans ce travail, nous présentons une nouvelle méthode permettant le calcul de similarités utilisant des arbres aléatoires. Cette approche étend celle des *Unsupervised Random Forest*, en utilisant des arbres extrêmement aléatoires. Notre approche est applicable à des jeux de données homogènes ou hétérogènes et possède des propriétés intéressantes que nous avons évalué empiriquement, telles que l'invariance aux transformations monotones des variables ou la résistance aux variables corrélées. Ces propriétés permettent de réduire les tâches de prétraitement. Une évaluation empirique de l'approche que nous proposons sur des jeux de données synthétiques et réels donne des résultats compétitifs vis-à-vis de la littérature.

Cependant, la méthode présente quelques inconvénients. Premièrement, bien que les expériences que nous avons réalisées nous donnent de bons résultats, nous n'avons pas à l'heure actuelle une définition claire des valeurs optimales pour les paramètres $n_{min}$ et $n_{trees}$ dans tous les cas. Deuxièmement, la taille des matrices de similarité, ainsi que la complexité de la méthode peuvent être problématique dans le cas de grands jeux de données.

La méthode que nous proposons ici peut être une bonne méthode candidate pour l'exploration de jeux de données, dans les cas où l'hétérogénéité des données ainsi que les tâches de prétraitement posent problème.

# Références

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Dalleau, K., M. Couceiro, et M. Smaïl-Tabbone (2018). Unsupervised extremely randomized trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 478–489. Springer.

Elghazel, H. et A. Aussem (2010). Feature selection for unsupervised learning using random cluster ensembles. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 168–175. IEEE.

Friedman, J. H. (2006). Recent advances in predictive (machine) learning. *Journal of classification 23*(2), 175–197.

Geurts, P., D. Ernst, et L. Wehenkel (2006). Extremely randomized trees. *Machine learning 63*(1), 3–42.

Shi, T. et S. Horvath (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics 15*(1), 118–138.

# Summary

In this paper we present a method to compute similarities on unlabeled data, based on extremely randomized trees. The main idea of our method, Unsupervised Extremely Randomized Trees (UET) is to randomly split the data in an iterative fashion until a stopping criterion is met, and to compute a similarity based on the co-occurrence of samples in the leaves of each generated tree. We evaluate our method on synthetic and real-world datasets by comparing the mean similarities between samples with the same label and the mean similarities between samples with different labels. Our empirical study shows that the method effectively gives distinct similarity values between samples belonging to different clusters, and gives indiscernible values when there is no cluster structure. We also assess some interesting properties such as invariance under monotone transformations of variables and robustness to correlated variables and noise. Finally, we performed hierarchical agglomerative clustering on synthetic and real-world homogeneous and heterogeneous datasets using UET. Our experiments show that the algorithm outperforms existing methods in some cases, and can reduce the amount of preprocessing needed with many real-world datasets.