Forêts aléatoires et clustering sous contrainte

Marie Gribouval[1,2], Simon Bernard[2], Laurent Heutte[2]
Albrecht Zimmermann[3], Bertrand Cuissart[3], Ronan Bureau[4]
[1] Master Science et Ingénierie des Données, Université de Rouen Normandie
[2] LITIS, [3] GREYC, [4] CERMN

Journée de l'axe Données Apprentissage Connaissances, 18 février 2021

RIN Tremplin SCHISM : Supporting chemoinformatics via interactive Unsupervised and semi-supervised data mining

- Coordinateur : Albrecht Zimmermann
- *Interactive Data Mining* :
  1. Exploration de motifs / Clustering
  2. Visualiser et analyser le résultat
  3. Fournir un retour (expert) sous la forme de containtes
  4. Ajuster le clustering
- Application : Chemoinformatics
  - Données : molécules
  - Objectif : interpréter les groupements de données en terme de propriétés physico-chimique ou de relation structure-activité
- WP2 (LITIS, GREYC, CERMN) :
  - Exprimer les données/contraintes en terme de dissimilarités
  - Forêts Aléatoires pour mesurer les dissmilarités
  - Sous-espaces de caractéristiques sous-jacents pour l'interprétabilité

Objectifs du stage

En amorce du WP2

1. Inférer un clustering à partir de forêts aléatoires (supervisé)
2. Proposer des outils d'analyse des groupements de molécules obtenus
3. Proposer un mécanisme d'intéraction avec l'expert (non-supervisé) :
   - Guider le retour d'expertise (interprétabilité)
   - Traduire ce retour en contraintes sur la mesure de dissimilarité
   - Adapter la mesure et le clustering résultant

Ce travail de stage porte principalement sur les points 1 et 2

Données

- Base de données de 1492 molécules
- Supervision : active ou non
- 2 types de descripteurs :
  1. Descripteurs topologiques (236 carac.) : propriétés physico-chimiques et topologiques des molécules
  2. Descripteurs *fingerprint* (2872 carac.) : présence ou absence de sous-structures (fragments) moléculaires (*sparse*)

CMPD_CHEMBLID	Activité	MW	AlogP	HBA	HBD	RB	HeavyAtomCount	...	path/walk 4 - Randic shape index	path/walk 5 - Randic shape index	reciprocal distance Randic-type index
CHEMBL3689722	1	339.3387	2.9054	3	1	3	25	...	0.17406803	0.08831989	3.77603611
CHEMBL3685101	1	321.3482	2.6999	3	1	3	24	...	0.17556562	0.08942457	3.73498365
CHEMBL3689631	1	419.3722	3.4533	4	2	5	30	...	0.16184006	0.08193867	4.11813294
CHEMBL3689554	1	351.3742	2.6779	4	2	4	26	...	0.17807532	0.08834177	3.8461213
CHEMBL3685056	1	318.3691	3.745	3	2	3	24	...	0.17192542	0.09465763	3.71964193
...	...	...	...	...	...	...	...	...	...	...	...
CHEMBL1290072	0	304.3425	3.9278	3	1	4	23	...	0.19506719	0.10966043	3.636137
CHEMBL228043	0	360.0014	2.6206	3	2	1	17	...	0.16468615	0.07020325	2.74705795
CHEMBL370283	0	375.6608	3.978	3	1	2	22	...	0.21819824	0.11737206	3.16220283
CHEMBL281470	0	276.2894	3.255	3	1	2	21	...	0.211469	0.11454945	3.33182497
CHEMBL3617729	0	191.2264	1.3048	2	0	2	14	...	0.19307463	0.08233775	2.62232957

1492

y

236

CMPD_CHEMBLID	0	16	32	17	3	1	136120670	203677720	...	-140434801	491535146	-557098051	-1303617740
CHEMBL3689722	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0
CHEMBL3685101	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0
CHEMBL3689631	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0
CHEMBL3689554	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0
CHEMBL3685056	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
CHEMBL1290072	1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	...	0.0	0.0	0.0	0.0
CHEMBL228043	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0
CHEMBL370283	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	...	1.0	0.0	0.0	0.0
CHEMBL281470	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	...	0.0	0.0	0.0	0.0
CHEMBL3617729	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	...	0.0	1.0	1.0	1.0

1492

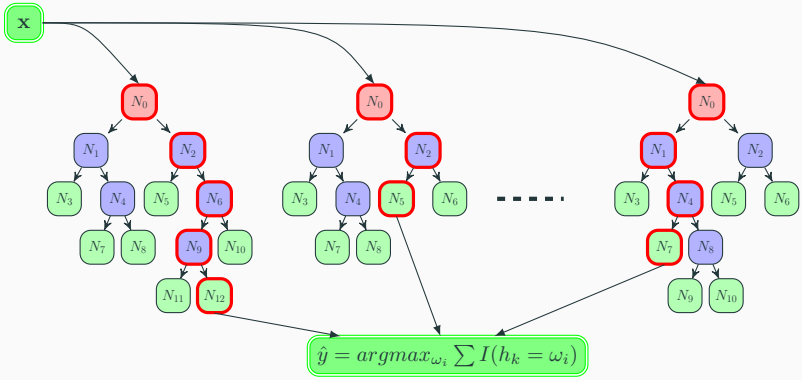
2873

Definition [Breiman 2001]

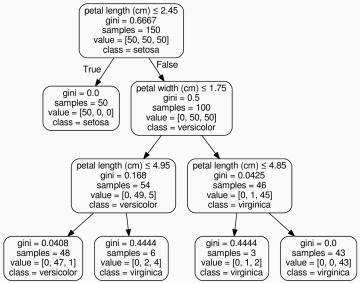
Une forêt aléatoire est un ensemble d'arbres de décisions, noté :

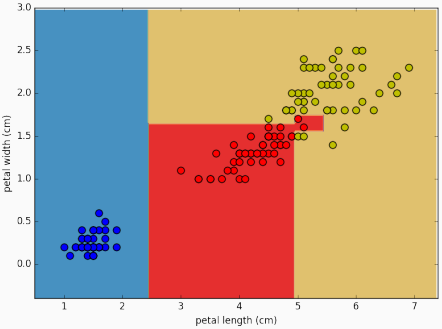
$$\{\ h_k = h(\mathbf{x}, \theta_k), \quad k = 1, ..., L\ \}$$

où les $\{\theta_k\}$ sont des vecteurs aléatoires i.i.d., et où chaque arbre h_k vote pour la classe à prédire pour une donnée $\mathbf{x}$ à classer.



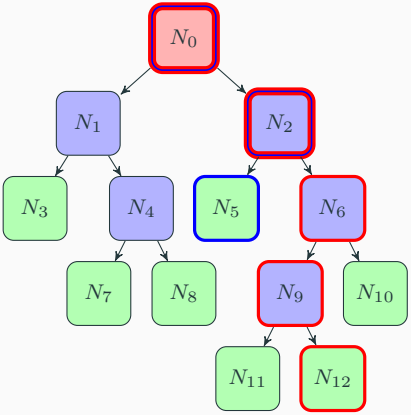
Un arbre de décision est naturellement un partitionnement des données





⇒ Une forêt aléatoire est un ensemble de *clustering*

Une forêt aléatoire permet également de mesurer des (dis)similarités entre instances [1]



- On note $l_k(\mathbf{x})$ la feuille de l'arbre h_k dans laquelle "tombe" $\mathbf{x}$ (ici $l_k(\mathbf{x}_i) = N_{12}$)
- La similarité $d^{(k)}(\mathbf{x}_i, \mathbf{x}_j)$ donnée par le k^{ieme} arbre est

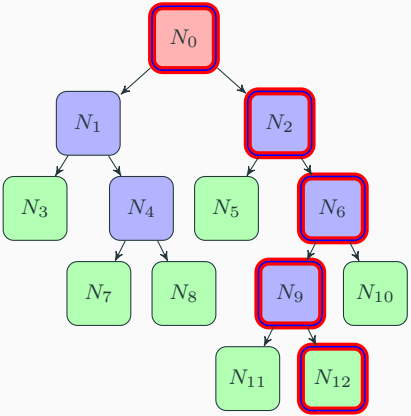
$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & si\ l_k(\mathbf{x}_i) = l_k(\mathbf{x}_j) \\ 0 & sinon \end{cases}$$

- Ici, $\mathbf{x}_i$ et $\mathbf{x}_j$ ne tombent pas dans la même feuille :

$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = 0$$

1. H. Cao, S. Bernard, R. Sabourin, L. Heutte, "Random forest dissimilarity based multi-view learning for Radiomics application", Pattern Recognition 88, 185-197, 2019

Une forêt aléatoire permet également de mesurer des (dis)similarités entre instances[1]



- On note $l_k(\mathbf{x})$ la feuille de l'arbre h_k dans laquelle "tombe" $\mathbf{x}$ (ici $l_k(\mathbf{x}_i) = N_{12}$)
- La similarité $d^{(k)}(\mathbf{x}_i, \mathbf{x}_j)$ donnée par le k^{ieme} arbre est

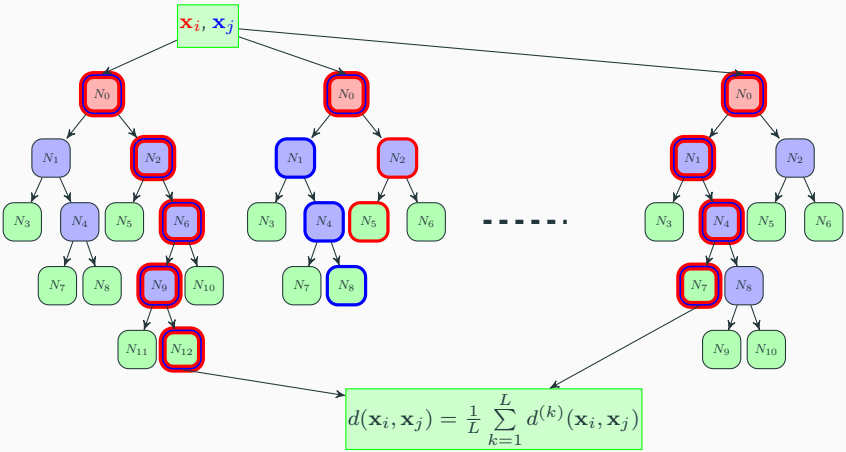
$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & si\ l_k(\mathbf{x}_i) = l_k(\mathbf{x}_j) \\ 0 & sinon \end{cases}$$

- Ici, $\mathbf{x}_i$ et $\mathbf{x}_j$ tombent dans la même feuille :

$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = 1$$

1. H. Cao, S. Bernard, R. Sabourin, L. Heutte, "Random forest dissimilarity based multi-view learning for Radiomics application", Pattern Recognition 88, 185-197, 2019

Une forêt aléatoire permet également de mesurer des (dis)similarités entre instances [1]



1. H. Cao, S. Bernard, R. Sabourin, L. Heutte, "Random forest dissimilarity based multi-view learning for Radiomics application", Pattern Recognition 88, 185-197, 2019

2 approches pour obtenir un clustering avec des forêts aléatoires

Cumulative Vote[2 3] : exploiter les partitionnements donnés par les arbres

- $\mathbf{U}^k$: matrice donnée par h_k, telle que l'élément de ligne i et de colonne j est :

$$\mathbf{U}^k_{i,j} = \begin{cases} 1 & si\ l_k(\mathbf{x}_i)\ est\ la\ j^{ieme}\ feuille \\ 0 & sinon \end{cases}$$

- Tri des $\mathbf{U}^k$ selon l'entropie de Shannon (descroissante)
- Mise à jour itérative :
  - On utilise la première matrice $\mathbf{U}^k$ comme référence (noté $\mathbf{U}^{(0)}$)
  - A l'itération i, on modifie $\mathbf{U}^{(0)}$ à l'aide de $\mathbf{U}^{(i)}$
  - En fin d'algorithme, $\mathbf{U}^{(0)}$ fournit des probabilités d'appartenance à chaque cluster
- Note : $\mathbf{U}^{(0)}$ contient toujours le même nombre de clusters

2. H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters", IEEE transactions on pattern analysis and machine intelligence, 30 (1), 160-173, 2007
3. F. Saeed, A. Ahmed, M. S. Shamsir, and N. Salim, "Weighted voting-based consensus clustering for chemical structure databases", Journal of computer-aided molecular design, 2(6), 675-684, 2014

2 approches pour obtenir un clustering avec des forêts aléatoires

Furthest Algorithm [4] : exploiter les dissimilarités

- $\mathbf{D}_H$: matrice donnée par la forêt, telle que l'élément de ligne i et de colonne j est $d(\mathbf{x}_i, \mathbf{x}_j)$
- Clustering itératif :
  1. Les 2 instances les plus dissimilaires : $\mathbf{x}_i, \mathbf{x}_j = \arg\max_{i,j} d(\mathbf{x}_i, \mathbf{x}_j)$
  2. $\mathbf{x}_i, \mathbf{x}_j$: centres de 2 clusters
  3. Chaque instance est affectée au cluster le plus "proche"
  4. $\mathbf{x}_k$ le plus dissimilaire $\mathbf{x}_i$ et $\mathbf{x}_j$: centre d'un nouveau cluster
  5. A chaque itération un coût est calculé :

$$d(\mathcal{C}) = \sum_{\mathbf{x}_i, \mathbf{x}_j : \mathcal{C}(\mathbf{x}_i) = \mathcal{C}(\mathbf{x}_j)} d(\mathbf{x}_i, \mathbf{x}_j) + \sum_{\mathbf{x}_i, \mathbf{x}_j : \mathcal{C}(\mathbf{x}_i) \neq \mathcal{C}(\mathbf{x}_j)} 1 - d(\mathbf{x}_i, \mathbf{x}_j)$$

     où $\mathcal{C}$ désigne le clustering actuel
  6. Tant que le coût diminue on ré-itère

4. A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation", ACM Transactionson Knowledge Discovery from Data (TKDD), vol. 1, no. 1, pp. 4–es, 2007.

Comment les hyperparamètres des forêts peuvent modifier les clustering résultant ?

Critères d'évaluation :

- Indice de Gini (↘)

$$G = \frac{1}{K}\sum_{i=1}^{K}\left(1 - \sum_{j=1}^{c}(\frac{n_j^{(i)}}{n^{(i)}})^2\right)$$

K : nombre de clusters, C : nombre de classes, $n_j^{(i)}$: nombre d'instances de la classe j dans le cluster i

- *Quality Partition Index* [5] (↗)

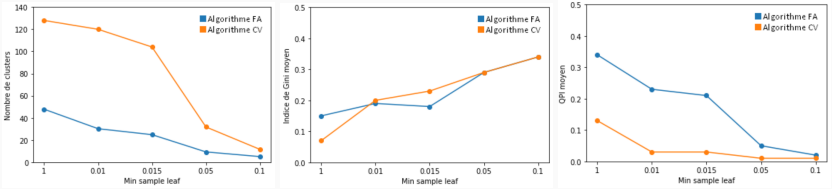
$$QPI = \frac{p}{p+q+r+s}$$

  - p : nombre de molécules actives dans les clusters actifs
  - q : nombre de molécules inactives dans les clusters actifs
  - r : nombre de molécules actives dans les clusters inactifs
  - s : nombre de singletons actifs

5. T. Varin, N. Saettel, J. Villain, A. Lesnard, F. Dauphin, R. Bureau, and S. Rault, "3d pharmacophore, hierarchical methods, and 5-ht4 receptor binding dat", Journal of enzyme inhibition and medicinal chemistry,23(5), 593-603, 2008

Comment les hyperparamètres des forêts peuvent modifier les clustering résultant ?

Influence de la profondeur des arbres sur les clustering :



- Arbres profonds : clusters nombreux, petits, homogènes. Analyse fine des descripteurs
- Arbres peu profond : clusters denses, éloignés, hétérogènes. Analyse "gros grains"

Influence du nombre d'arbres sur les clustering :

- CV : plus il y a d'arbres dans la forêt, mieux c'est (sans surprise).
- FA : résultats plus hétéroclites. Analyse à approfondir.