Clustering de molécules à partir de données quantitatives

Martin Jolif

October 2023

1 Jeu de données

Pour faire ce clustering, j'ai utilisé le jeu de données téléchargeable à cette adresse : `https://github.com/deepchem/deepchem/blob/master/datasets/delaney-processed.csv`

Une visualisation de celui ci ci-dessous :

	Compound ID	ESOL predicted log solubility in mols per litre	Minimum Degree	Molecular Weight	Number of H-Bond Donors	Number of Rings	Number of Rotatable Bonds	Polar Surface Area	measured log solubility in mols per litre
0	Amigdalin	-0.974	1	457.43200000000013	7	3	7	202.32	-0.77
1	Fenfuram	-2.885	1	201.225	1	2	2	42.24	-3.3
2	citral	-2.579	1	152.237	0	0	4	17.07	-2.06
3	Picene	-6.6179999999999986	2	278.35400000000004	0	5	0	0.0	-7.87
4	Thiophene	-2.232	2	84.14299999999999	0	1	0	0.0	-1.33
5	benzothiazole	-2.733	2	135.191	0	2	0	12.89	-1.5
6	2,2,4,6,6'-PCB	-6.545	1	326.437	0	2	1	0.0	-7.32
7	Estradiol	-4.138	1	272.388	2	4	0	40.46	-5.03
8	Dieldrin	-4.533	1	380.913	0	5	0	12.53	-6.29
9	Rotenone	-5.246	1	394.4230000000002	0	5	3	63.22	-4.42
10	2-pyrrolidone	0.243	1	85.10600000000002	1	1	0	29.1	1.07

Figure 1: Jeu de données sur les molécules

2 Clustering

Après avoir centré et normalisé les données, j'ai appliqué différents algorithmes de clustering au jeu de données. Pour essayer d'obtenir de bons résultats, j'ai essayé d'optimiser les paramètres des différents algorithmes, par exemple en effectuant la méthode du coude pour l'algorithme K-means (voir le code).

Enfin pour pouvoir comparer les résultats des différents algorithmes j'ai utiliser les index *Silhouhette Coefficient*, *Calinski-Harabasz Index* et le *Davies Bouldin Index*. Finalement, j'ai obtenu les résultats suivant :

	SC	CH	DB	nombres de clusters	eps	min_samples	bandwith	damping
K-Means	0.319307	370.755062	1.074502	5.0	NaN	NaN	NaN	NaN
DBSCAN	0.281903	120.236217	1.754494	2.0	1.2	18.0	NaN	NaN
Birch	0.277991	275.104806	1.163664	5.0	NaN	NaN	NaN	NaN
Mean shift	0.348766	83.349734	0.755817	6.0	NaN	NaN	3.2	NaN
OPTICS	-0.104407	91.297366	1.483337	5.0	1.2	18.0	NaN	NaN
Affinity Propagation	0.37267	166.358215	0.977793	3.0	NaN	NaN	NaN	0.91

Figure 2: Résultats des différents algorithmes de clustering sur le jeu de données

3 Code Informatique

Tout le code informtique nécessaire à l'obtention de ces résultats est disponible à l'adresse suivante : `https://github.com/martinjolif/molecules-clustering` dans le fichier Molecules Clustering on quantitative data.

Mon code s'est inspiré du tutoriel à l'adresse suivante : `https://www.kaggle.com/code/maximgolovatchev/unsupervised-learning-clustering-tutorial#Clustering-Methods`