Année universitaire 2024–2025

Fouille de données olfactives : clustering de molécule odorantes par GCL (Graph Constrastive Learning)

Projet de [√] « Recherche » [√] « Recherche et développement » [ ] « Développement »

Encadrement :
Fabrice Guillet, LS2N-DUKe, Polytech Nantes, Fabrice.Guillet@univ-nantes.fr
Angélique Villière, GEPEA-Flaveur, Oniris, Angelique.Villiere@oniris-nantes.fr

Cadre de l'étude

L'équipe « Flaveur » du laboratoire de biochimie GPEA à Oniris est une équipe phare dans l'étude de l'arôme des aliments par olfactométrie. L'olfactométrie combine l'analyse chimique et la perception humaine d'un juge entraîné (un *nez*). Elle permet, après une extraction de l'arôme d'un aliment, d'analyser et d'identifier, les molécules odorantes constituant cet arôme. (voir vidéo https://images.cnrs.fr/video/2819 et https://lejournal.cnrs.fr/articles/de-la-molecule-a-lodeur). Une des principales difficultés rencontrées réside dans la forte variabilité des capacités de perception des odeurs par les juges (physiologique, anosmie, culturel, représentation mentale des odeurs).

Objectif

L'objectif de ce travail est d'apprendre un modèle de représentation vectorielle des odeurs émises par des molécules à l'aide de technique d'apprentissage automatique par réseau neuronal profond (deep learning).
En remarquant que chaque molécule peut être représentée par un graphe, le travail consistera à appliquer des méthodes de *Graph Neural Network* (GNN, clustering à base de graphes) sur les données

Pour cela, 4 informations complémentaires peuvent être croisées :

1° les données des sites *the good scent company* et *flavor net* qui répertorient les associations molécule-odeurs communes sur de nombreux produits.
2° des *données expérimentales* en faible volume (small data), récoltées à Oniris, où des juges identifient (par un nom) les odeurs détectées sur des molécules.
3° les données sur les propriétés physico-chimiques des molécules (sites *PubChem* et *NIST*) : fonction ester…
4° un *graphe de connaissances* (représentation sémantique, hiérarchie d'odeurs) disponible dans SketchOscent, où les odeurs sont regroupées en familles (pôles, super-pôles, etc …), avec l'hypothèse qu'au sein d'un même pôle les odeurs se ressemblent davantage (similitude intra pôle élevée) qu'entre 2 pôles distincts (similitude inter pôle faible).

En remarquant que chaque molécule peut être représentée par un graphe, le travail consistera à appliquer des méthodes de *Graph Neural Network* (GNN, clustering à base de graphes) sur les données (1,2,3) afin d'extraire les sous graphes fréquents caractéristiques de chaque odeur, chaque sous graphe correspondant à la partie odorante de la molécule (forme de la clé odorante dans un modèle clé/serrure).
En complément, on pourra construire des sous-ensembles d'odeurs similaires et de les comparer à ceux proposés par le graphe de connaissance de SketchOScent (données 5).

Travail à réaliser

Le travail consistera à :
1) lire les articles *[1] Sisson, L. (2023)*, et *[2] Sanchez-Lengeling et al. (2019)*, *Wiltschko et al. (2018)* (voir la section *références à lire*)
2) comprendre les modèles et méthodes mises en œuvres : clustering, representation learning, GCN, GAE
3) reproduire le code (disponible dans un git) et les résultats de l'article [1]
4) apporter une 1ere amélioration des résultats obtenus en remplaçant le GCN par un GCL tel que décrit dans l'article [3]
5) apporter une 2eme amélioration en utilisant des données supplémentaires sur les fonctions chimiques des molécules connues pour être liées aux odeurs (ex : fonction ester, cétones, aldéhydes, lactones, etc…).

Tous les résultats obtenus devront être testés et évalués sur les données, et également comparés aux autres méthodes disponibles (issues de l'état de l'art)

Les clusters résultant de ces méthodes seront analysés et comparés aux pôles du graphe de connaissances des odeurs. Pour cela il sera nécessaire de définir une similarité entre odeurs au sein du graphe des odeurs (dite distance sémantique dans le graphe), et de la comparer à la similarité entre odeur issue du clustering (distance sur les données) ; mais aussi d'évaluer la qualité du clustering (silhouette, calinsky, dunn, modularity, clustering coefficient).

Références à lire

[1] Sisson, L. (2023). Olfactory Label Prediction on aroma-chemical Pairs. *arXiv preprint arXiv:2312.16124*. https://arxiv.org/html/2312.16124v2

[2] Sanchez-Lengeling, B., Wei, J. N., Lee, B. K., Gerkin, R. C., Aspuru-Guzik, A., & Wiltschko, A. B. (2019). Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*. https://arxiv.org/abs/1910.10685 and also http://ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html

[3] Ren, Y., Liu, B., Huang, C., Dai, P., Bo, L., & Zhang, J. (2019). Heterogeneous deep graph infomax. *arXiv preprint arXiv:1911.08538*. https://arxiv.org/pdf/1911.08538

[4] *R. Ying*, A. Wang*, J. You, J. Leskovec, (2020). Frequent Subgraph Mining by Walking in Order Embedding Space.* (voir http://snap.stanford.edu/frequent-subgraph-mining/ et http://snap.stanford.edu/class/cs224w-2021/slides/12-motifs.pdf)

Compétences

Langage : R, python (scikit, pytorch/tensorflow/keras)
Fouille de données : méthodes non supervisées, deep learning , GNN
Autres : python rdflib, gitlab, notebooks