Graph Neural Network for chemistry

Léon Ter[*]
University of Nantes, France
leon.ter@etu.univ-nantes.fr

Duc Mai Chu[*]
University of Nantes, France
duc.chu@etu.univ-nantes.fr

1 Sharing of the work

Léon Ter: Analysis of the MoSS results, experiment GCN, training and analysis on Sirtuin6

Duc Mai Chu: Test on MoSS with the three different input files, find the Sirtuin6 dataset

2 Frequent subgraph mining for molecules

To test out the MoSS (Molecular Subsstructure miner) program, we run 3 different examples of input files in the SMILES format.

- example1.dat: 6 entries
- example2.dat: 3 entries
- steroids.dat: 17 entries

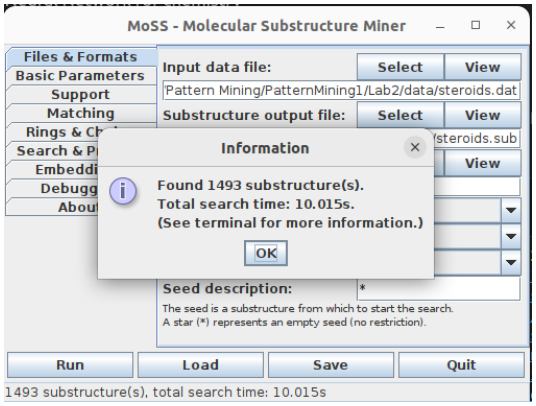

Figure 1: Information tab after finding the substructures of the steroids.dat file

[*]Equal contribution.

Ter et al., Graph Neural Network for chemistry. *Proceedings of the Second Learning on Graphs Conference (LoG 2023)*, PMLR 231, Virtual Event, November 27–30, 2023.

```
search statistics:
maximum search tree height   : 25
number of search tree nodes  : 660762
number of created fragments  : 1295079
number of created embeddings : 10256186
insufficient support pruning : 486783
perfect extension pruning    : 39219
equivalent sibling pruning   : 0
canonical form pruning       : 101799
ring order pruning           : 0
duplicate fragment pruning   : 0
non-closed fragments         : 659269
fragments with open rings    : 0
fragments with invalid chains: 0
auxiliary invalid fragments  : 0
accesses to repository       : 0
comparisons with fragments   : 0
actual isomorphism tests     : 0
comparisons with embeddings  : 0
```

Figure 2: Statistics after finding the substructures in the terminal

The substructure output file contains the following elements:

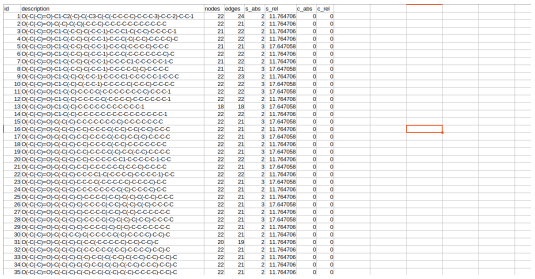

Figure 3: Sample of the result of the steroids.dat

We can separate the columns into two categories, which are the key measures for evaluation: support metric (Starting with 's' for the substructure in the focus part of the database: s_abs absolute support, s_rel relative support. Starting with 'c' for the substructure in the complement part of the database) and structure attributes (node, edges)

We can observe the most frequent sequences in the dataset by ordering on s_rel.

id	description	nodes	edges	s_abs	s_rel	c_abs	c_rel
1228	O-C	2	1	17	100	0	0
1289	C12(-C)-C-C-C(-C(-C-C-C)-C-2-C-C-C-1)-C	14	15	17	100	0	0
847	O-C1-C-C-C(-C-1(-C)-C-C-C-C)-C-C	13	13	15	88.23529	0	0
852	O-C1-C-C-C-C-1(-C)-C-C-C(-C-C-C-C)-C	15	15	15	88.23529	0	0
858	O-C(-C-C)-C(-C)-C-C)-C-C(-C-C-C)-C-C	15	14	15	88.23529	0	0
866	O-C(-C-C)-C(-C)(-C-C-C-C)-C-C-C-C-C	15	14	15	88.23529	0	0
876	O-C(-C-C)-C(-C)(-C)-C-C(-C-C)-C(-C)-C	14	13	15	88.23529	0	0
890	O-C(-C)-C(-C)(-C-C)-C(-C)-C(-C-C-C)-C-C	15	14	15	88.23529	0	0
899	O-C(-C)-C(-C)(-C-C-C-C)-C(-C)-C-C-C-C	15	14	15	88.23529	0	0

Figure 4: Result of steroids.dat sorted by support

Among the two substructures found the most, we have id1228 and id1289. While looking over the different substructures, it appears normal that the simple chain of oxygen (O) and carbon (C) id1228 is one of the most frequent among the steroids dataset. The length of the chain isn't a criteria to tell how frequent a substructure is, as the second most is a quite complex substructure and others like id634 are much less frequent. (support 64)

3 Graph neural networks for molecules (GCN)

The experiments for this section are contained in the python notebook gcn_molecule.ipynb. Just to try out the basic concepts of the library we run first the same code as in the tutorial to make sure it

works. We then work on the Sirtuin6 Small Molecules dataset, for classification, which includes 100 molecules with descriptors to determine the candidate inhibitors of a target protein. The molecules are grouped based on low- and high-BFE which we use for the classification. It can be accessed with the following link: https://archive.ics.uci.edu/dataset/748/Sirtuin6+small+molecules-1

We perform evaluation on the regular measures for models with the function classification_report from sklearn.metrics. It covers precision, recall, f1-score and support for the classes.

- Precision : measures the accuracy of positive predictions.
- Recall : measures the ability of the model to find all positive samples.
- f1-score : mean of precision and recall.
- suport : number of true samples in each class.

	precision	recall	f1-score	support
0	0.70	0.88	0.78	8
1	0.90	0.75	0.82	12
accuracy			0.80	20
macro avg	0.80	0.81	0.80	20
weighted avg	0.82	0.80	0.80	20

Figure 5: Result of classification report

The accuracy is evidence that the model correctly classifies with 80% of the samples.

Observing precision, recall and the f1_score allows us to individually assess the quality of the classification over each of the possible values predicted. We notice High_BFE (class 1) has a higher precision rate that Low_BFE (class 0) while invertly High_BFE has a lower recall rate. The precision suggests that the model incorrectly classifies low_BFE as high_BFE more often. On the contrary the recall suggests there are fewer false negatives for low_BFE.

We can also print the confusion matrix which compares the predicted labels with the true labels.

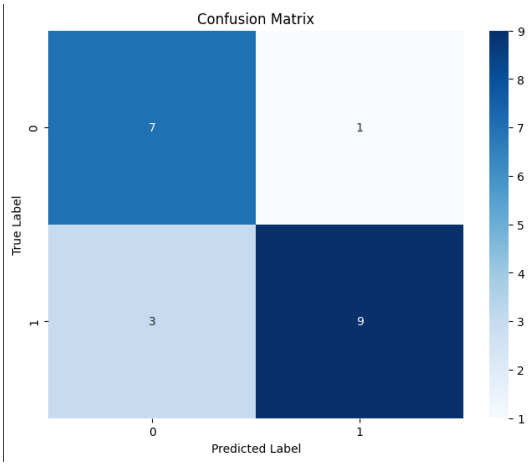

Figure 6: Result of the confusion matrix

To read the confusion matrix in this case, we could look at the false positives (1). The false positives are especially important in this context of drug molecules where we want to minimize the number of times we wrongfully classify a molecule as a class it doesn't belong.