Tox21_AR

Information

Total 9362 molecules (graphs) in the dataset.

5617 molecules (graphs) in the training set.

2808 molecules (graphs) in the testing set.

937 molecules (graphs) in the scoring set.

The molecules are represented as graphs (with neighboring matrices), instead of Molecular Formula.

Files

1. <train/test/score>_labels.csv

A csv file which contains the graph index and corresponding label, where 0 means non-toxic and 1 means toxic.

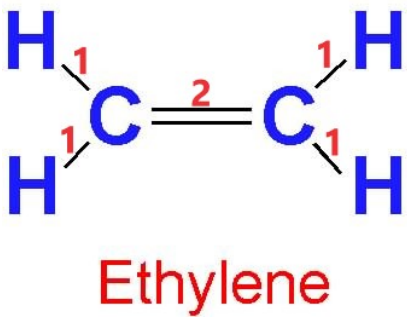
2. <train/test/score>_graph_size.csv

A csv file which contains the graph index and graph size (the number of nodes in this graph).

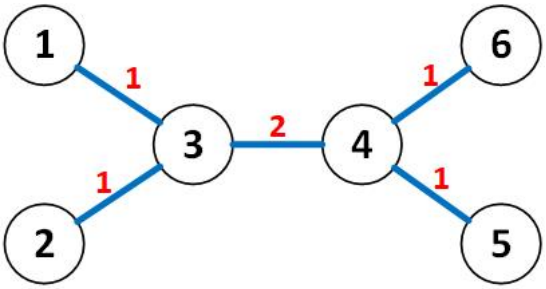
3. <train/test/score>_graphs.csv

A csv file which contains the neighboring matrices of graphs in <train/test/score> set. The largest molecule (graph) has 132 atoms (nodes). So each graph in this file is represented as a 132x132 matrix. There are three types of edges in these graphs. If two nodes are not connected, the value in the matrix is 0. If they are connected by TYPE-1 edge, the corresponding value in the neighboring matrix is 1. If they are connected by TYPE-2 edge, the corresponding value in the neighboring matrix is 2. And if connected by TYPE-3 edge, the value is 3.

Here is an example: The molecule formulation of Ethylene is C2H4. The structure is:



There are 2 types of edges in this molecule. The corresponding graph representation is:



The matrix representation is:

	1	2	3	4	5	6	...	...	132
1	0	0	1	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0
3	1	1	0	2	0	0	0	0	0
4	0	0	2	0	1	1	0	0	0
5	0	0	0	1	0	0	0	0	0
6	0	0	0	1	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0
132	0	0	0	0	0	0	0	0	0

Each graph is represented as a 132x132 matrix. In this example there are only 6 nodes, therefore only the first 6 rows and the first 6 columns have non-zero values. If two nodes are not connected, the corresponding value is 0. If they are connected, the value represents the edge type. The matrix representation is symmetrical, i.e. the graph is an undirected graph.

4. <train/test/score>_nodes.csv

Each line in this csv file is the feature (atomic feature) of the node in the graph. By default, each graph has 132 nodes. If the number of the nodes in one graph is less then 132, i.e. N <= 132, then the features of the (132-N) nodes are set as 0. Features of the N nodes are their real features.

DNN

1. You can try to use traditional neural networks (e.g. TensorFlow MNIST example code) to solve this problem. Maybe you only need the neighboring matrix.
2. You can try some new and popular neural networks, e.g. Graph Convolutional Neural Networks. You can use the node features provided in the dataset.
3. You can implement any data processing (preprocessing or postprocessing) algorithms in your code.

ATTENTION

1. One person per project.
2. Auto-submission will be provided.
3. Repeated submission allowed.
4. Python 3.
5. TensorFlow 1.5.0 and Numpy only.

6. Stick to the output format and the file names specified.