Graph-Based Molecule Classification using Graph Edit Distance and KNN

May 26, 2025

Objective

The goal of this work was to build a machine learning model that classifies molecules as either **active** or **inactive** against HIV, based on their structural graph representation.

Each molecule is represented as a graph, where:

- **Nodes** represent atoms (with chemical symbols as labels)
- **Edges** represent undirected covalent bonds between atoms

Methodology

We approached the problem as a graph classification task using the following steps:

1. **Graph Construction**
   GXL files were parsed using `xml.etree.ElementTree`. Graphs were constructed using the `NetworkX` library, where each node stored its atom type (e.g., 'C', 'O', etc.).
2. **Similarity Computation**
   For each pair of graphs, we computed the **Graph Edit Distance (GED)**. GED was approximated using bipartite graph matching based on mismatches in node symbols.
3. **Feature Extraction**
   Each molecule was represented as a feature vector of GEDs to all training molecules.
4. **Classification**
   We used **K-Nearest Neighbors (KNN)** with $k = 5$, training the model using the GED-based feature vectors.

Validation Results

The model achieved the following performance:

- **Validation Accuracy: 0.98** (98%)

This high accuracy shows the model's strong ability to generalize on unseen molecules in the validation set.

Test Predictions

The test set contained 1,500 molecules without known labels. After running predictions, we saved the output in a file named `test.tsv` with the following format:

```
5600    inactive
5641    inactive
5650    inactive
5651    inactive
5665    active
21643   active
21684   active
21686   active
```

Tools and Libraries

- Python 3.10 (compatible with GraKeL)
- NetworkX
- Scikit-learn
- NumPy, SciPy
- xml.etree.ElementTree

Conclusion

Our graph-based KNN model using GED achieved a validation accuracy of 0.98, demonstrating that structural similarity is a powerful feature for classifying molecular activity. The approach is transparent, explainable, and effective for small to medium-sized molecular graphs.