Toxicity Prediction of Chemical Compounds

Recchi Giovanni, Matteo Manias, Simone Cargnin, Gabriele Scognamiglio

Abstract

Most people are exposed to numerous chemicals through food, cleaning products, and medicines, some of which can be toxic. This project investigates the use of Graph Neural Networks (GNNs) for predicting the toxicity of chemical compounds, leveraging the TOX21 dataset. Our approach involves preprocessing the dataset, extracting molecular features, and training multiple GNN architectures. The results demonstrate that GNN models outperform traditional machine learning techniques in prioritizing ROC AUC. Key challenges include handling imbalanced data and optimizing model hyperparameters to minimize false negatives, ensuring safety-critical predictions.

1 Introduction

Exposure to chemical compounds is an inevitable aspect of daily life, yet some of these substances may pose significant toxicological risks. Predicting the toxicity of chemical compounds is vital for public health and safety. This project explores leveraging advanced machine learning models, particularly Graph Neural Networks (GNNs), to improve the accuracy of toxicity predictions. Using the TOX21 dataset, we compare the performance of GNNs against classical machine learning methods.

2 Related Work

The TOX21 dataset has been widely utilized for predictive toxicology tasks. Prior studies have demonstrated the effectiveness of deep learning methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in extracting meaningful patterns from chemical data. Recent advancements in GNNs provide a novel approach by explicitly modeling the graph structure of molecular data, yielding promising results in tasks such as molecular property prediction.

3 Proposed Method

Our experimental pipeline consists of:

- **Data Analysis**: Exploring the dataset to identify key features and distributions.
- **Data Preprocessing**: Cleaning and transforming molecular features extracted from SMILES notation.
- **Feature Extraction**: Employing 801 dense features for chemical descriptors and 272,776 sparse features for chemical substructures.
- **Model Training**: Training and tuning various GNN architectures, including Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), NNConv, and GCN with edge convolution.
- **Performance Evaluation**: Benchmarking against traditional machine learning models, such as Random Forest and Gradient Boosting, using metrics like ROC AUC.

4 Dataset and Benchmark

The TOX21 dataset comprises 12,000 training samples and 650 test samples representing chemical compounds. Each sample includes 12 binary labels indicating active/inactive outcomes. The dataset is highly imbalanced, with significantly more inactive compounds, which poses challenges for maintaining recall without compromising precision.

5 Experimental Results

The experimental results demonstrate the superior performance of Graph Neural Networks (GNNs) compared to classical machine learning models. Key findings include:

- **Baseline Models:** Random Forest and Gradient Boosting established performance benchmarks but struggled with recall due to dataset imbalance.
- **GNN Models:** GCN and GAT outperformed baseline models, achieving higher ROC AUC scores. NNConv demonstrated exceptional recall, critical for minimizing false negatives.
- **Threshold Optimization:** Adjusting classification thresholds improved the trade-off between precision and recall, ensuring critical toxic compounds were not overlooked.
- **Impact of Data Imbalance:** Stratified sampling and weighting during training partially mitigated challenges posed by the imbalance, enhancing the overall robustness of the models.

These results validate the potential of GNNs in effectively modeling the relationships within chemical data for toxicity prediction.

6 Conclusion and Future Work

This study highlights the effectiveness of GNNs for toxicity prediction, achieving superior results compared to classical models. Key contributions include the successful application of GNN architectures, mitigation of data imbalance challenges, and optimization of classification thresholds for safety-critical applications.

One notable finding is the trade-off between recall and precision. While achieving high recall, which is critical for identifying toxic compounds and minimizing false negatives, the precision of the model decreases significantly. Despite this, the GNN approach still outperforms traditional methods, representing a substantial step forward in predictive toxicology.

Future work will focus on:

- Expanding the dataset to improve model generalization.
- Exploring advanced data balancing techniques such as synthetic data generation.
- Investigating additional GNN architectures for improved performance.
- Developing interpretability tools to provide insights into model decisions.

In conclusion, the advancements presented in this project pave the way for more accurate and reliable toxicity prediction methodologies, marking a meaningful progression in the field despite the challenges in achieving an optimal balance between recall and precision.

References

1. TOX21 dataset: https://paperswithcode.com/dataset/tox21-1
2. Project GitHub Repository: https://github.com/RezaCDoobary/DrugDiscovery-Tox21
3. Comprehensive paper on TOX21: https://arxiv.org/pdf/2208.04852
4. Model-specific insights: https://arxiv.org/pdf/2209.05582