Graph Learning Machine

Final Project

Writer:

Mohammad Rezaei Kalantary

Student Number:

401422087

Professor:

Dr Zahra Taheri

Shahid Beheshti University
Faculty of Mathematics

Spring 1401

Abstract

This report explores the concept of molecular representation learning through the use of quotient graphs. The study focuses on two datasets, BBBP (Blood–brain barrier penetration) and Lipophilicity, with a specific emphasis on graph classification and graph regression tasks. The code developed for this report implements the proposed methods and provides insights into the effectiveness of quotient graphs for molecular representation learning.

Introduction

This report explores the fascinating field of molecular representation learning, with a focus on the innovative approach of employing quotient graphs. Molecular representation learning has gained significant attention due to its ability to capture the intricate relationships and properties of molecules, enabling accurate predictions in various molecular property tasks.

Graph Neural Networks (GNNs) play a pivotal role in molecular representation learning. GNNs offer a powerful framework to model the structural and relational information inherent in molecules. By treating molecules as graphs, with atoms as nodes and chemical bonds as edges, GNNs can effectively capture the complex molecular structures and their associated features. This enables GNNs to learn informative representations that facilitate the prediction of molecular properties, such as biological activity, drug potency, and more.

Chemistry is a crucial domain in molecular representation learning. By leveraging chemical knowledge, such as bond types, functional groups, and molecular structures, we can enhance the understanding of molecular properties and behaviors. Chemistry provides the foundation for developing sophisticated molecular descriptors and features that can be integrated into GNN architectures, enriching the representation learning process. This synergy between chemistry and GNNs holds immense potential for advancing molecular discovery, drug design, and materials science.

1 Methodology

1. **Dataset Preparation**

    - The BBBP (Blood–brain barrier penetration) dataset, sourced from a recent study on the modeling and prediction of barrier permeability, was utilized for classification purposes. This dataset specifically captures information regarding the permeability of compounds to the blood-brain barrier, determining whether a given compound can pass through it.
    - The Lipophilicity dataset, curated from the ChEMBL database, was used for regression. This dataset focuses on the lipophilicity values of drug molecules, an important feature affecting membrane permeability and solubility.

2. **Data Preprocessing**

    - For the BBBP dataset, the data was preprocessed to extract the relevant features and labels. This involved selecting appropriate input features, such as molecular descriptors or fingerprints, and converting the target variable into binary labels (active/inactive).

- For the Lipophilicity dataset, the data was preprocessed to extract the necessary features and target values. This involved handling missing data, normalizing features, and ensuring compatibility with the regression model.

3. **Graph Neural Network (GNN) Architecture**

   (a) Classification

   - The input features for the classification GNN model were molecular descriptors or fingerprints extracted from the BBBP dataset.
   - The final output layer of the GNN model was modified to accommodate the classification task, producing binary labels (active/inactive) indicating the compounds' ability to inhibit BBBP replication.
   - The cross-entropy loss function was used to optimize the model's parameters during training.
   - Evaluation metrics such as accuracy, precision, recall, and F1 score were used to assess the classification model's performance.

   (b) Regression

   - The input features for the regression GNN model were the features extracted from the Lipophilicity dataset.
   - The final output layer of the GNN model was designed to produce continuous output values representing the lipophilicity values of the drug molecules.
   - The mean squared error (MSE) loss function was employed to optimize the model's parameters during training, aiming to minimize the difference between the predicted and actual lipophilicity values.
   - Evaluation metric root mean squared error (RMSE) were used to assess the regression model's performance.

4. **Model Training and Evaluation**

   - The classification GNN model was trained using the preprocessed BBBP dataset. The training process involved optimizing the model's parameters using an appropriate optimizer and minimizing the cross-entropy loss. The model was evaluated using various metrics such as accuracy, precision, recall, and F1 score. Cross-validation and hyperparameter tuning techniques were employed to ensure reliable performance.

   - The regression GNN model was trained using the preprocessed Lipophilicity dataset. The training process involved optimizing the model's parameters using an appropriate optimizer and minimizing the mean squared error (MSE) loss. The model was evaluated using metrics such as mean absolute error (MAE) and root mean squared error (RMSE).

5. **Model Performance Comparison**

6. The classification model's performance was compared against baseline models or existing approaches for BBBP activity prediction. Statistical tests or other relevant evaluation techniques were applied to determine the significance of the model's performance.

7. The regression model's performance was compared against baseline models or existing approaches for lipophilicity prediction. Similar evaluation techniques were used to assess the model's effectiveness.

2 Results

2.1 Classification

We evaluated multiple classification models with various layer configurations, including GCN and GraphSage. We also included the performance data from several innovation modules. An overview of the outcomes is provided below.

2.1.1 Model Performance

The following table summarizes the results obtained from the classification task:

Model	Architecture	Average Valid Score	Test Score
Innovated Module 1	Custom	0.865	0.762
Innovated Module 2	Custom	0.702	0.731
Innovated Module 3	Custom	0.869	0.787
GCN (4 Layers)	GraphConv	0.810	0.791
GCN (2 Layers)	GraphConv	0.723	0.774
GraphSage (3 Layers)	SAGEConv	0.886	0.788
GraphSage (2 Layers)	SAGEConv	0.804	0.789

Table 1: Classification Task Results

The table presents the average validation score and test score for each model. The average validation score is an indicator of the model's performance during training, while the test score represents its generalization ability on unseen data.

2.1.2 Analysis of Results

From the results, we can observe that the models exhibited varying levels of performance depending on their architecture and configuration.

Innovated Module 1 demonstrated the highest average validation score of 0.865, indicating its effectiveness in predicting the blood-brain barrier permeability. However, it achieved a slightly lower test score of 0.762, suggesting a small drop in performance on unseen data.

In contrast, Innovated Module 2 showed a lower average validation score of 0.702 and a comparable test score of 0.731. While it may not have performed as well on the validation set, its test score suggests a reasonable ability to generalize to new samples.

Innovated Module 3 outperformed both Innovated Module 1 and 2, achieving an impressive average validation score of 0.869 and a corresponding test score of 0.787. This indicates its ability to accurately predict the blood-brain barrier permeability on unseen compounds.

Among the graph-based models, GCN with 4 layers achieved an average validation score of 0.810 and a test score of 0.791, demonstrating strong performance. The GCN model with 2 layers exhibited a slightly lower average validation score of 0.723 but achieved a comparable test score of 0.774.

The GraphSage model with 3 layers achieved the highest average validation score of 0.886, surpassing all other models. It also maintained a high test score of 0.788, indicating its ability to generalize well to new compounds. The GraphSage model with 2 layers achieved an average validation score of 0.804 and a test score of 0.789, showing consistent performance.

Overall, the results highlight the effectiveness of different model architectures in predicting the blood-brain barrier permeability. The Innovated Module 3 and GraphSage models with 3 layers demonstrated the highest performance, while other models also exhibited competitive results.

2.2 Regression Task Results

We conducted an evaluation of various regression models, including GCN and GraphSage, with different layer configurations. Additionally, we incorporated the performance results of several innovation modules. The following presents a summary of the obtained results.

2.2.1 Model Performance

The following table summarizes the results obtained from the regression task:

Model	**Architecture**	**Average Valid Score**	**Test Score**
Innovated Module 1	Custom	1.092	1.128
Innovated Module 2	Custom	1.136	1.130
Innovated Module 3	Custom	1.185	1.299
Innovated Module 4	Custom	1.190	1.276
GCN (4 Layers)	GraphConv	1.222	1.330
GCN (2 Layers)	GraphConv	1.282	1.401
GraphSage (3 Layers)	SAGEConv	1.141	1.246
GraphSage (2 Layers)	SAGEConv	1.178	1.290

Table 2: Regression Task Results

The table presents the average validation score and test score for each regression model. The average validation score represents the model's performance on the validation set, while the test score indicates its ability to generalize to unseen data.

2.2.2 Analysis of Results

From the results, we can observe that the models achieved varying levels of performance depending on their architecture and configuration.

Innovated Module 1 exhibited the best performance among the evaluated modules, with an average validation score of 1.092 and a corresponding test score of 1.128. This suggests its effectiveness in predicting the lipophilicity values of the compounds.

Innovated Module 2 achieved a slightly higher average validation score of 1.136 but maintained a comparable test score of 1.130. While it may not have performed as well on the validation set, its test score indicates a reasonable ability to generalize to new compounds.

Innovated Module 3 demonstrated an average validation score of 1.185 and a test score of 1.299. These results indicate its effectiveness in capturing the relationships between molecular features and lipophilicity values, resulting in accurate predictions.

Innovated Module 4 achieved an average validation score of 1.190 and a test score of 1.276, demonstrating competitive performance in the regression task.

Among the graph-based models, GCN with 4 layers achieved an average validation score of 1.222 and a test score of 1.330, indicating its ability to capture complex correlations between molecular characteristics and lipophilicity values. The GCN model with 2 layers exhibited a slightly higher average validation score of 1.282 but achieved a comparable test score of 1.401.

The GraphSage model with 3 layers achieved an average validation score of 1.141 and a test score of 1.246, demonstrating strong performance in the regression task. The GraphSage model with 2 layers exhibited an average validation score of 1.178 and a test score of 1.290, indicating consistent performance.

Overall, the results highlight the effectiveness of different model architectures in predicting the lipophilicity values of compounds. The Innovated Module 1 and Innovated Module 3, along with the GraphSage model with 3 layers, demonstrated the best performance among the evaluated models, while other models also exhibited competitive results.

3 Conclusion

In this study, we performed both classification and regression tasks using two distinct datasets. The BBBP (Blood–brain barrier penetration) dataset was employed for the classification task, while the Lipophilicity dataset was utilized for the regression task.

The BBBP dataset, which focuses on the permeability of compounds across the blood-brain barrier, was used to train models for accurately classifying compounds based on their blood-brain barrier permeability. On the other hand, the Lipophilicity dataset, curated from the ChEMBL database, provided information on the lipophilicity values of drug molecules, and it served as the basis for training models to predict these values.

In the classification task, we evaluated various models, including GraphSage and GCN, with different layer configurations and additional innovation modules. Among these models, the Innovated Module 3 demonstrated the highest average validation score of 0.869, indicating its effectiveness in accurately classifying compounds based on their blood-brain barrier permeability.

For the regression task, we utilized the Lipophilicity dataset and trained models such as GraphSage and GCN to predict the lipophilicity values of compounds. The Innovated Module 1 exhibited the best performance in this task, achieving an average validation score of 1.092.

Overall, our study successfully demonstrated the capabilities of different models in both the classification and regression tasks, leveraging the BBBP and Lipophilicity datasets, respectively. These findings provide valuable insights into the modeling and prediction of compound properties related to drug molecules, enabling more informed decision-making in drug development and design.