Graph Classification using Frequent Subgraph Mining

Yearning For The Machines

February 2025

1 Feature Engineering

The primary approach for feature extraction in this study involves the use of frequent subgraphs as discriminative features. The features were selected using the χ^2 (Chi-square) statistical test, which evaluates the importance of each frequent subgraph in distinguishing between graph classes.

2 Frequent Subgraph Mining

To extract meaningful features, we first mined subgraphs using the gSpan algorithm. gSpan efficiently discovers frequent subgraphs in a dataset without candidate generation, making it suitable for large-scale graph mining. The mined subgraphs serve as the basis for constructing feature representations.

3 Feature Construction

Feature vectors were constructed in two key steps, implemented in the scripts `identify.py` and `convert.py`.

3.1 Identifying Discriminative Subgraphs

The script `identify.py` processes the dataset to identify frequent subgraphs. It:

- Runs the gSpan algorithm on the training set to extract subgraph patterns.
- Evaluates the support of each subgraph in the dataset.
- Selects the most discriminative subgraphs using the χ^2 test, ensuring only the most informative patterns are retained.

These subgraphs are then used as the primary features for classification.

3.2 Feature Encoding

The script `convert.py` encodes graphs into feature vectors:

- Each graph is checked for the presence or absence of the selected frequent subgraphs.
- A binary feature vector is constructed where each entry represents whether a specific subgraph appears in the graph.
- The resulting vectors are stored as NumPy arrays for efficient processing by the classifier.

This transformation enables the application of machine learning models to classify graphs based on their subgraph structures.

4 Classification and Results

The extracted features were used to train a machine learning classifier. The classification model was evaluated on a test set, achieving an ROC-AUC score of 73%. This result demonstrates that frequent subgraph-based features provide meaningful signals for distinguishing graph classes.

5 Conclusion

Frequent subgraph mining, combined with χ^2-based feature selection, effectively captures structural patterns in graphs for classification. The approach leverages gSpan for subgraph discovery and converts graphs into structured feature vectors, enabling the use of standard machine learning techniques. Future work could explore alternative subgraph selection methods or deep learning-based approaches for improved performance.