Question No 3 Report

Team Who Cares

February 2025

1 Problem Statement

We have been assigned to extract discriminative features from given graph classification datasets without using any classifier or neural network approach.

2 Our Method

2.1 Core Algorithm

Our method follows these steps:

1. **Dataset Splitting**: Split the given dataset into training and testing sets.
2. **Frequent Subgraph Mining**: Identify frequent subgraphs based on dynamically determined parameters:
    - Support threshold
    - Minimum number of vertices
    - Maximum number of vertices
3. **Dynamic Parameter Selection**: The support and vertex constraints are chosen adaptively based on the size and characteristics of the dataset.
4. **Subgraph Extraction**: Apply the Gaston algorithm to the training dataset using the selected parameters to extract frequent subgraphs.
5. **Feature Construction**:
    - If the number of extracted frequent subgraphs is fewer than 100, construct train and test features based on the presence or absence of each subgraph in the corresponding graph.
    - If more than 100 subgraphs are found, apply a *filtering mechanism** (described below) to select the 100 most relevant subgraphs for feature construction.

2.2 Filtering Mechanism

For filtering, we experimented with a method that:

1. **Purely discriminative feature construction**: Select only those subgraphs which are present in one class and absent in another. This allows easy class discrimination.

2. **Score-based filtering**: Of all the remaining subgraphs, filter based on a **score function** that ranks the remaining subgraphs. We select '80 - number_of_purely_discriminative_subgraphs' by picking the top subgraphs. The scoring functions used are:

   - **Discriminative score**:

     $$ds = |n_{pos} - n_{neg}| \tag{1}$$

     where n_{pos} is the number of positive class graphs containing the subgraph, and n_{neg} is the number of negative class graphs containing the subgraph.

   - **Mutual Information**:

     $$MI(S) = \sum_{x \in 0,1} \sum_{y \in 0,1} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{2}$$

     This measures how much uncertainty is reduced when a particular subgraph is present.

3. **Total coverage-based selection**: We find all graphs that do not contain any of the subgraphs selected so far. We then select additional subgraphs from these uncovered graphs until we reach 100 features or the number of uncovered graphs becomes zero.

2.3 Experiments and Results

Out of many experiments that we did, we present the results of a few below.

Table 1: Experiments and their results

Dataset	**Support**	**MinV**[1]	**MaxV**[2]	**Train ROC**	**Test ROC**
Mutagenicity	0.1	1	7	0.91	0.85
NCI-H23	0.1	1	4	0.92	0.82
NCI-H23	0.2	1	4	0.85	0.77

2.3.1 Observations

- The discriminative score and mutual infomration methods provide similar ranking results

- Mutagenicity data is much easier to discriminate, and also the number of samples being less, it is easier to come up with features
- Even without using very basic binary features like presence or absence of triangle, whether average degree is greater than 2 or not, we get a decent baseline.

3 References

- Siegfried Nijssen, Joost N. Kok *A Quickstart in Structure Mining Can Make a Difference*
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar *Introduction to Data Mining*
- Saiful Islam, University of Buffalo *A Structural Feature-Based Approach for Comprehensive Graph Classification*
- For code refinement, report refinement, and grammatical helps *ChatGPT, Perplexiy*