Data Augmentation Based on Null Model for Graph Classification

Qi Xuan,¹ Zeyu Wang,¹ Jinhuan Wang,1* Yalu Shan,¹

¹ Institute of Cyberspace Security, College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China
xuanqi@zjut.edu.cn, Vincent.Wang@outlook.com, jhwang@zjut.edu.cn, yalushan@foxmail.com

Abstract

In network science, null model is typically used to generate a series of graphs based on randomization under certain condition, which is widely used as a term of comparison to verify whether the networks in question display some non-trivial features, such as community structure. Since such non-trivial features may play a significant role in graph classification, the null model could provide a new perspective for regularization, so as to lead to the enhancement of classification performance. In this paper, we propose a novel data augmentation framework based on null model for graph classification, which contains four parts: feature ranking, graph data augmentation, data filtration, and model retraining. Moreover, in this framework, three heuristic null model generation methods are proposed for different features. Experiments are conducted on five famous benchmark datasets, and the results show that our framework has promising performance, providing a new direction of data augmentation for graph classification.

Introduction

Data augmentation is a technology that artificially expands the training datasets by allowing limited data to generate more equivalent data. Research in the computer vision field has been relatively mature, but that in network science is still in its infancy. What’s more, more researches in network science, especially in graph classification, mainly focused on the graph structure and proposed some heuristic methods from the perspective of nodes or communities to modify the graph topology structure. Such a method can be considered as a way to generate augmented data by introducing a tiny disturbance into the original data and then altering the parameter through retraining the model, which also can be thought of as a regularization method but lacks the interpretability of the data augmentation.

Graph classification aims to identify the category labels of graphs in a dataset by using features that are extracted by handcraft, graph kernels, graph embedding, or graph neural networks (Xuan et al. 2021). Some of the non-trivial features of the graph are mainly focused on by the classifier. Therefore, one should pay more attention to these non-trivial features when proposing data enhancement methods, which can not only improve the classifier performance, but also provide the interpretability of the graph classification.

This phenomenon is similar to the principle of the null model. Null models are pattern-generating models that deliberately exclude a mechanism of interest, and as a popular analysis tool are applied to the dynamics of complex networks mainly depending on its ability to explore the non-trivial features of the graph. Null models are mainly applied in ecological and biogeographic data, and they are used in comparisons to quantify complex network properties such as community structure (Newman 2006), microbial diversification (Zhai et al. 2018), assortativity (Pastor-Satorras, Vázquez, and Vespignani 2001), degree correlation (Mahadevan et al. 2006), epidemic spreading rate (Estrada et al. 2016), routing efficiency (Nian and Fu 2014), pattern detection (Ulrich and Gotelli 2013), microbial diversification (Zhai et al. 2018), etc. Therefore, in network science, null models are especially notable because they reveal important network properties that could not be directly quantified in some situations.

In this paper, we propose a novel data augmentation method based on different order null models, containing four components: features ranking, null model graph augmentation, data filtration, and model retraining. The idea is to find out the non-trivial features of the graph and develop the null model according to the non-trivial features to produce more virtual data for retraining the graph classifier. Since the virtual data we generated mainly guarantees that the non-trivial features will not change, their availability remains to be verified. In view of this, we adopt the data filter proposed in Zhou et al. (2020) to filter fine augmented examples from the generated data. We demonstrate that the method can achieve a significant improvement of performance on graph classification and have strong interpretability for graph classification. Specifically, we have the following contributions:

• We extend the null model to the graph classification task and propose a null model-based data augmentation framework, which contains features ranking, null model graph augmentation, data filtration, and model retraining for the first time.

• We explore the interpretability of the graph classification, which can promote the further development of the graph classification task, by the null model which is used to verify the non-trivial features displayed by the graph.
• We develop three specific null models for non-trivial features, i.e., percentage of leaf node (P), average betweenness centrality (BC) and average closeness centrality (CC) which play a significant role in the graph classification.

• We apply our data augmentation framework based on four different feature extraction methods into graph classification, and the experimental results on five real-world network datasets demonstrate the effectiveness of our framework.

The rest of paper is organized as follows. First, in Sec. 2, we review the related works on null model and data augmentation of graph classification. Then, in Sec. 3, we introduce a null model-based data augmentation framework and three (approximate) null model generation methods. Thereafter, in Sec. 4, we present extensive experiments with detailed discussions. Finally, in Sec. 5, we conclude the paper and outline future work.

Related Work

Null Model

Null model is a pattern-generating model that is based on randomization of ecological data or random sampling from a known or imagined distribution (Gotelli and Graves 1996). Null model as an analysis tool is applied in ecological and biogeographic scenes in the past few decades, such as the laws of species migration, island rules and the spatial pattern of trees in temperate forests (Biddick and Burns 2021 Carrer et al. 2018 Reimann et al. 2019). Different from other generating models, the null model constructs a model that deliberately excludes a mechanism being tested (Gotelli 2001). Researchers have used null models to simulate this process and found the rules, which fully shows that the null model can explain the internal mechanism of the model. More precisely, the null model consists of a network that can be seen as one specific graph with some structural attributes but that is otherwise taken to be a random network instance. For undirected graphs, some studies have proposed several null model generation methods, such as 0-order, 1-order, 2-order, 2.5-order null model graphs (Gjoka, Kurant, and Markopoulou 2013 Mahadevan et al. 2007 2006 Newman Strogatz, and Watts 2001 Chung and Lu 2002b a). Such null models are very helpful for us to explore the nature of modeling and structures of complex networks. It is worth mentioning that However, there is no study to analyze the graph classification task by utilizing the null models. In this paper, we adopt four null model with different orders and three novel null models generated according to different non-trivial features to enhance the performance of graph classification.

Data Augmentation

Data augmentation is a common method to solve the problems that caused by insufficient data set or model overfitting, which is widely applied in computer vision (Cubuk et al. 2019 DeVries and Taylor 2017) and natural language processing (Fadaee, Bisazza, and Monz 2017 Sahin and Steedman 2019). So far, data augmentation in graph data is in its infancy. Zhou et al. (Zhou et al. 2020) developed three heuristic methods to generate the virtual data and successfully yields an average improvement accuracy on graph classification tasks. In node classification tasks, Zhao et al. (Zhao et al. 2020) discovered that neural edge predictors can effectively encode class-homophilic structure to promote intra-class edges and demote inter-class edges in given graph structure and they leveraged these insights to improve performance in GNN-based node classification via edge prediction. Dong et al. (Dong et al. 2020) propagated the labels of the training set through the graph structure and expand the training set. Although the above methods all improve the classifier performance by expanding the data set, they all lack interpretability, which is the researches are most concerned about. In this paper, we proposed a data augmentation framework based on null models, which can provide the explanatory basis for the enhancement of graph classification.

Methodology

In this section, we first state the problem of data augmentation on graph classification. And then, we introduce our optimization framework based on null model which could be utilized to improve the accuracy of the graph classification task. Finally, we present series of heuristic null model-based data augmentation strategies.

Notations and Problem Statement

As we all know, the graph classification task is aimed to divide the graphs into their correct categories. In this paper, we only discuss undirected and unweighted graphs. Generally, a graph dataset that consists of several graphs can be represented as \( D = \{ (G_i, y_i) | i = 1, 2, \cdots, N \} \), where \( G_i \) is an undirected and unweighted graph, and \( y_i \in Y \) is the label of graph \( G_i \). Given a graph \( G = (V, E) \), where \( V = \{ v_1, v_2, \cdots, v_n \} \) denotes the set of nodes, and \( E \subseteq (V \times V) \) denotes the set of edges.

For a given graph dataset \( D \), we can split it into training set \( D_{train} \), validation set \( D_{val} \) and test set \( D_{test} \). The training set and validation set are used to train a classifier \( C \) in advance, called the original classifier. And then, we can expand the training set by utilizing data augmentation method and get the augmented graphs \( D_{aug} \). Further, we union the training graphs and the augmented graphs to retrain the classifier and obtain a new classifier \( C' \). Note that, the augmented data is expressed as \( D'_{train} \), and the labels of the augmented graphs are the same as the original training graphs. Finally, the test data is used to test the accuracy of the original classifier and the new classifier. Specifically, the purpose of data enhancement is to optimize the classifier and enhance the efficiency of graph classification.

Targeted Optimization Framework

Specific optimization process, that is, a targeted data augmentation method for different datasets to optimize the classifier, contains feature ranking, graph augmentation, data filtration, and model retraining in top-down. As shown in Figure, the dataset will be firstly split as the training set
Algorithm 1: Targeted Optimization Framework

**Input:** Dataset $D$, number of iteration $T$

**Parameter:** The ratio of augmentation

**Output:** Optimization classifier $C'$

1: Split the dataset as $D_{train}$, $D_{val}$, $D_{test}$;
2: features Ranking, get the targeted augmentation model $\mathcal{F}$;
3: Pre-train $D_{train}$, $D_{test}$, get the original classifier $C$;
4: Initialize iteration $= 0$;
5: while iteration $< T$ do
6: Graph augmentation: $D_{aug} = \mathcal{F}(D_{train})$;
7: Data filtration: $D'_{train} = \text{filter}(D_{aug})$;
8: $D_{train} \leftarrow \text{append}(D'_{train})$;
9: Optimize the classifier $C : C' \leftarrow \text{retrain}(C, D_{train})$;
10: iteration $= \text{iteration} + 1$;
11: $C \leftarrow C'$;
12: end while
13: return $C'$;

$D_{train}$, the validation set $D_{val}$ and the test set $D_{test}$. And then, $D_{train}$ and $D_{val}$ will be used to pre-train the classifier and get the original classifier $C$. Meanwhile, the topological features extracted on training set are selected by feature ranking mechanism to obtain the targeted features. Next, we design the graph augmentation model with the targeted features, which can be adopted to augment the training set $D_{train}$ to obtain the augmentation set $D_{aug}$. Finally, $D_{aug}$ will be feed into the data filter module to get the valid augmented set $D'_{train}$. Using $D'_{train}$ and $D_{train}$ to re-train the classifier $C$ can obtain the optimization classifier $C'$. The Figure 1 describes the procedure of the specific optimization framework and the pseudocode is shown in Algorithm 1.

**Feature Ranking**

Feature ranking is an important module of the augmentation framework, which aims to find the key features of the dataset. Here, we will give the detailed introduction and the details are as follows.

**Graph features.** In order to find the significant features of graph data in graph classification as much as possible, we adopt eleven graph features, *Number of nodes* ($N$), *Number of edges* ($E$), *Average degree* ($D$), *Percentage of leaf nodes* ($P_L$), *Largest eigenvalue of the adjacent matrix* ($EV$), *Network density* ($DS$), *Average clustering coefficient* ($C$) (Borgatti and Everett 1997), *Lind, Gonzalez, and Herrmann* 2005, *Average betweenness centrality* ($BC$) (Wolfe 1995), *Average closeness centrality* ($CC$) (Bavelas 1950, Beauchamp 1965), *Average eigenvector centrality* ($EC$) (Bonacich 1972, 2005), *Average neighbor degree of graph* ($ND$), for subsequent feature selection.

**Features Selection.** In order to find the important feature accurately, we utilize the *Gini importance* computed by the *Random Forest* (Breiman 2001) classifier to evaluate the feature importance. As shown in Figure 2, the colors describe the values of *Gini importance* of eleven features on different datasets. The feature with the high *Gini importance* (darker square) plays a more significant role than others in graph classification. With the completion of random forest training, we can get the *Gini* index set $S_{Gini}$ on the importance of features. The *Gini* index set $S_{Gini}$ can be denoted as

$$S_{Gini} = \{S_i | i = 1, 2, \cdots, m\},$$

where $m$ is the number of the kinds of features, $S_i$ represents the importance of the $i$th feature. And the feature select function $\mathcal{F}_{Max}$ will deal the *Gini importance* index set and return the key feature $f_{key}$. The key feature is denoted as

$$f_{key} = \mathcal{F}_{Max}(S_{Gini}).$$

**Graph Data Augmentation**

Here, we use four naive null models to design the graph augmentation module, including 0-order (0k), 1-order (1k), 2-order (2k), and 2.5-order (2.5k) null model. Moreover, we also construct three heuristic null model-based graph augmentation modules, Leaf Node Augmentation (LNA), Betweenness Centrality Augmentation (ADA-BC), and Closing Centrality Augmentation (ADA-CC).

**Null Model Augmentation.** As we all know, there are two methods to generate the null model. One is based on the configuration model (Ren et al. 2017), and the other is based...
on rewiring edges. In this paper, the method of rewiring edges is used to construct 0k, 1k, and 2.5k null models. The null model with different orders has different properties to keep consistent with the original network. Among them, the 0k null model is the same as the original graph in terms of the number of nodes and average degree. The 1k null model maintains the consistency of node degree distribution with the original graph on the basis of 0k null model. The 2k null model has the same joint degree distribution as the original graph. The 2.5k null model is the average clustering coefficient with the same joint degree distribution and degree correlation as the original graph.

**Leaf Nodes Augmentation (LNA).** LNA is an augmentation strategy based on the percentage of leaf nodes. Figure 3 (a)-(f) describe the whole process of generating LNA null model. For a given graph $G = (V, E)$, we can selectively rewiring the edges connected with leaf nodes in the graph to ensure that the percentage of leaf nodes is the same as that of the original graph. Firstly, it is easy to get the leaf node set $V_L$ as follows:

$$V_L = \{v_i | \text{Degree}(v_i) = 1, i = 1, 2, \ldots, n\},$$

where the $n$ is the number of leaf nodes. And the edges which connect to leaf node is defined as

$$E_L = \{(V_L(i), V_L(i).\text{neighbor}) | i = 1, 2, \ldots, n\}.$$  

Then, on the condition that the neighbor node of the leaf node cannot become a new leaf node after disconnecting the edge, the edge set is filtered to obtain the candidate edge set $E'_L$, and corresponding leaf node set $V'_L$. Next, one can get the set of leaf nodes to be changed $V_{LC}$ randomly:

$$V_{LC} = \{v_i | i = 1, 2, \ldots, n * \alpha\} \in V'_L,$$  

and the set of edges removed from $G$ could be defined as

$$E_{del} = \{(V_{LC}(i), V_{LC}(i).\text{neighbor}) | i = 1, 2, \ldots, n * \alpha\} \in E'_L,$$  

where $\alpha$ is the budget of edge modification and $\alpha \in [0, 1]$.

Because the higher degree of a node, the more significant it is, we choose the node which the two-hop node is a leaf node and has the highest degree to connect. After select the node $v \in V_{LC}$, one can choose the node with the highest degree of two-hop nodes of each leaf node and form a set $V_{LT}$. And then, we can get the set of edges:

$$E_{add} = \{(V_{LC}(i), V_{LT}(i)) | i = 1, 2, \ldots, n\},$$

where $n$ is the length of $V_{LC}$. Finally, after the addition and deletion of the connection, we can get a new graph $G'$:

$$G' = (V, (E \cup E_{add}) \setminus E_{del}).$$  

**Approximate Data Augmentation (ADA).** Many graph features are closely related to the global topology of graph, and even the local feature of single node could be affected by all other nodes of the graph, such as the betweenness centrality and closeness centrality. This makes it difficult to keep some features consistent with the original graph in the process of changing the network structure of the graph. Therefore, when we do specific feature augmentation, we use the feature approximation method to reduce the impact on the feature in the augmentation process. Here, we apply the concept of feature approximation to the specific augmentation of average betweenness centrality and average closeness centrality. We replace these two global features with $f$, replace betweenness centrality and closeness centrality with $f$, which are discussed below.

As show in the Figure 4 for a given graph $G = (V, E)$, we can randomly select an edge $e \in E$ and treat it as a potential deleted edge. And if there is no leaf node in this edge, it can become a removable edge $e_{del} = (v_a, v_b)$. In the process of generating null model, if $f_{value}(v_a) \geq f_{value}(v_b)$, we will keep the degree of the node $v_a$ to be invariant and plan to connect a node $v_c$ where $f_{value}(v_c)$ is approximate to $f_{value}(v_b)$. And if the reconnected edge is not in the original graph, it will become an new edge, denoted as $e_{add} = (v_a, v_c)$. That’s it for one edge change, and it will happen several times in the process of graph augmentation. At the same time, we set iteration parameter $T$ for this augmentation method, and choose the best-augmented graph to return. In a sense, as long as $T$ is large enough, we can get an augmented graph with high similarity.
Data Filtration

As we all know, data augmentation is widely used in image recognition and classification. There are many means of image data augmentation, such as changing pixels, rotating images, and geometric transformation, etc. Although data augmentation changes the image, the new image is easy to judge by human. In other words, it retains some significant features in the augmentation process. The data augmentation changes the image, the new image is easy to judge by human. In other words, it retains some significant features in the augmentation process. The graph data is different from the image, after data augmentation, the graph data may lose some important information of the original graph. If the original label is given to the graph obtained by simple data augmentation, it may not only fail to optimize the classifier but will be counterproductive. Therefore, this paper uses a data filtering mechanism to filter the enhanced data.

The filtering mechanism utilizes the label credibility $r$ and threshold of the graph $\theta$ to filter. We know that when the graph $(G_i, y_i)$ is input into the classifier, we can get a prediction matrix $p \in \mathbb{R}^{Y \times 1}$, which represents a probability distribution that the graph is divided into incorrect category, where $|Y|$ represents the number of classes of the graph. Then we can get an average probability matrix $q_k$ as follows:

$$q_k = [q_{k1}, q_{k2}, \ldots, q_{k|Y|}]^T = \frac{\sum_{y_i = k} p_i}{\Omega_k}, \quad (9)$$

where $k$ represents the category, $q_{kj}$ represents the probability that the classifier classifies the $i$th class graph into the $j$th class, and $\Omega_k$ represents the number of $k$th class graphs.

The label reliability of a graph can be defined as:

$$r_i = p_i^T q_{y_i}. \quad (10)$$

And we can know that the more likely the graph to be correctly classified will have a larger label probability.

The threshold $\theta$ is defined as:

$$\theta = \arg\min_\theta \sum_{(G_i, y_i) \in D_{test}} \Phi[(\theta - r_i) \cdot g(G_i, y_i)], \quad (11)$$

where $g(G_i, y_i) = 1$ if $C(G_i) = y_i$ and $g(G_i, y_i) = -1$ otherwise, and $\Phi(x) = 1$ if $x>0$ and $\Phi(x) = 0$ otherwise.

Experiments and Results

In this section, we conduct some experiments to validate the effectiveness of our graph data augmentation framework on graph classification task. We first introduce the datasets, followed by the feature extraction methods, and the experiment setting. After that, we show the experimental results with discussion.

Datasets

In order to access the effectiveness of our framework, we adopt five commonly used benchmark datasets in experiments including BZR(Jeffrey et al. 2004), BZR_MD(Jeffrey et al. 2004), 2011, 2012, COX2(Jeffrey et al. 2004), MUTAG(2011, 2012) and OHSU. Among them, BZR, BZR_MD, COX2 and OHSU are biological networks and MUTAG is chemical networks. The statistics of these datasets are summarized in Table 2.

Feature Extraction Methods

We utilize four graph classification methods such as SF, NetLS, gl2vec, and Graph2vec, where the SF and Graph2vec are graph embedding, and the rest are graph kernel models.

- SF (de Lara and Pineau 2018) is an embedding method which relies on spectral features of the graph. And it performs graph classification based on the spectral decomposition of graph Laplacian.
- NetLS (Tsitsulin et al. 2018) is a graph kernel model which compares graphs and achieves graph classification by extracting a compact signature that inherits the formal properties of the Laplacian spectrum.
- gl2vec (Tu et al. 2019) constructs vectors for feature representation which is generated by static or temporal network graphlet distributions and a null model to compare them with random graphs. It is mainly used for network classification in static and time-oriented networks.
- Graph2vec (Narayanan et al. 2017) is the first unsupervised embedding approach for an entire network, which can learn data-driven distributed representations of arbitrary sized graphs.

Experiments Settings

Dataset split. The dataset is divided into the training set, validation set, and test set in the ratio of 7:1:2. And we use 5-fold cross validation, that is, randomly split the dataset 25 times. Parameters setting. We set the random state of the Random Forest classifier to 0. Meanwhile, we set the modified edge connection ratio of each augmented model to 0.2
Table 1: Graph classification results of original and augmented models. "-" represents 2.5k null model has a higher limitation for BZR_MD, which leads to the virtual data is not enough for the experiment. The best results are marked in bold.

| Datasets | Aug Model | st | svm-n | logistic | kNN | rf | svm-n | logistic | kNN | rf | svm-n | logistic | kNN | rf | Avg. Gain Ratio |
|----------|-----------|----|-------|---------|----|----|-------|---------|----|----|-------|---------|----|----|----------------|
| BZR      | original  | 0.796| 0.74 | 0.805 | 0.838 | 0.807| 0.722 | 0.802 | 0.814 | 0.795| 0.819 | 0.843 | 0.837 | 0.827 | 0.852 | 0.846 | 2.40% |
|          | 1k        | 0.799| 0.811 | 0.808 | 0.835 | 0.804| 0.81 | 0.813 | 0.833 | 0.804| 0.84 | 0.858 | 0.833 | 0.824 | 0.855 | 0.842 | 2.71% |
|          | 2k        | 0.805| 0.804 | 0.816 | 0.841 | 0.816| 0.813 | 0.811 | 0.833 | 0.804| 0.84 | 0.858 | 0.833 | 0.824 | 0.855 | 0.842 | 2.71% |
|          | 2.5k      | 0.803| 0.809 | 0.808 | 0.842 | 0.813| 0.807 | 0.812 | 0.83 | 0.803| 0.84 | 0.86 | 0.834 | 0.848 | 0.843 | 2.95% |
|          | LNA       | 0.806| 0.809 | 0.815 | 0.842 | 0.815| 0.814 | 0.815 | 0.832 | 0.802| 0.822 | 0.855 | 0.835 | 0.824 | 0.889 | 0.849 | 2.55% |
|          | ADA-B-C   | 0.803| 0.802 | 0.818 | 0.843 | 0.809| 0.808 | 0.813 | 0.83 | 0.801| 0.83 | 0.852 | 0.844 | 0.799 | 0.824 | 0.843 | 2.16% |
|          | ADA-CC    | 0.803| 0.804 | 0.808 | 0.843 | 0.812| 0.809 | 0.813 | 0.829 | 0.803| 0.83 | 0.852 | 0.844 | 0.799 | 0.824 | 0.843 | 2.16% |
| BZR_MD   | original  | 0.603| 0.569 | 0.613 | 0.608 | 0.633| 0.595 | 0.611 | 0.529 | 0.589| 0.571 | 0.561 | 0.584 | 0.591 | 0.557 | 0.567 | 2.97% |
|          | 0k        | 0.607| 0.611 | 0.646 | 0.622 | 0.633| 0.604 | 0.635 | 0.601 | 0.596| 0.594 | 0.606 | 0.622 | 0.596 | 0.606 | 0.607 | 2.97% |
|          | 1k        | 0.611| 0.621 | 0.64 | 0.619 | 0.627| 0.633 | 0.627 | 0.581 | 0.591| 0.62 | 0.605 | 0.622 | 0.596 | 0.606 | 0.607 | 2.97% |
|          | 2k        | 0.614| 0.627 | 0.633 | 0.619 | 0.627| 0.643 | 0.625 | 0.594 | 0.596| 0.624 | 0.598 | 0.605 | 0.572 | 0.61 | 0.61 | 5.15% |
|          | LNA       | 0.612| 0.611 | 0.633 | 0.619 | 0.628| 0.639 | 0.626 | 0.595 | 0.582| 0.616 | 0.588 | 0.615 | 0.582 | 0.618 | 0.599 | 6.17% |
|          | ADA-B-C   | 0.612| 0.619 | 0.637 | 0.625 | 0.637| 0.64 | 0.643 | 0.593 | 0.612| 0.618 | 0.603 | 0.629 | 0.581 | 0.589 | 0.604 | 6.58% |
|          | ADA-CC    | 0.612| 0.607 | 0.648 | 0.625 | 0.64 | 0.63 | 0.625 | 0.569 | 0.614| 0.618 | 0.606 | 0.638 | 0.58 | 0.604 | 0.599 | 5.14% |

Table 2: Statistics of datasets used in experiments.

| Datasets | No.Graphs | No.Classes | Avg. Nodes | Avg. Edges |
|----------|-----------|------------|------------|------------|
| BZR      | 405       | 2          | 43.75      | 38.44      |
| BZR_MD   | 306       | 2          | 21.3       | 225.06     |
| COX2     | 467       | 2          | 41.22      | 43.45      |
| MUTAG    | 188       | 2          | 17.93      | 43.79      |
| OHSU     | 79        | 2          | 82.01      | 439.66     |

and set the number of the iteration of approximate augmentation as 5.

In order to verify the benefit of our framework, we adopt four feature extraction methods combined with four classical classifiers for graph representation and classification, and take the average result of 25 times to eliminate the randomness of the experiments. In other words, each group will be trained by the 16 graph classification frameworks composed of four classification models and four embedding models, and calculate the average classification accuracy gain. Among them, for all kernel and embedding methods, we set the embedding dimension to 128. In addition, we set the number of iterations $T$ to 5.

### Comparison and Evaluation

In order to facilitate comparison and validation, we trained all datasets corresponding to each enhancement model. Table 1 shows the comparison results between the original model and the various augmentation models mentioned in Section 3. We define the gain coefficient $C_{gain}$ as:

$$C_{gain} = \frac{Acc_{aug} - Acc_{ori}}{Acc_{ori}} \times 100\%$$

where $Acc_{aug}$ denotes the augmented classification accuracy, and $Acc_{ori}$ denotes the accuracy of the original classifier. In the following, we will evaluate the augmentation effect of the null model and targeted augmentation framework respectively.

### Analysis of Null Model

To evaluate the performance gain by the null model, we set the original classifier as baseline and utilize the classifier with null model augmentation to training the five datasets. It can be seen from Table 1 that null model augmentation has a significant gain on graph classification. Also, from the observation, we can conclude that it does not apply to all datasets that the higher the order, the better the augmentation effect. While the difference between the different orders null model is the features limitation, it is the main reason that resulted above phenomenon. Then, we propose the targeted augmentation.

### Analysis of Specific Augmentation

According to the analysis of null model augmentation, we propose the targeted features data augmentation based on the null model. When we augment a dataset, we only select an important
feature besides the number of edges and the number of nodes to augment. And Figure 2 reports the result of the features ranking of the five datasets. Based on the ranking results in Figure 2, we can find that: 1) EV and EC are more important than other features on BZR, BZR_MD, COX2, and MUTAG; 2) BC is more important than other features in COX2; 3) In OHSU dataset, P accounts for a larger proportion than other features. Due to the difficulty of EV and EC, we choose P, BC, and CC for specific augmentation and set the 0k null model as the baseline. Based on this setting, we do experiments on five datasets and list the results in Table 1.

From Table 1, we can analyze different augmentation models respectively:

- **LNA.** Compared with the original classifier and the 0k null model, LNA can see a more significant gain on each dataset, especially in the OHSU dataset, because P is more important on the OHSU dataset than that on other datasets. The augmentation strategy ensures that P is constant, its performance is slightly better than for the 0k null model on the rest four datasets.

- **ADA-BC & ADA-CC.** For the augmentation effect of BC and CC, we can see that they are up and down relative to the 0k null model augmentation on different datasets. The main reason for this phenomenon is that the data augmentation of BC and CC is not a null model, which ensure the feature is constant. It is only ensure the feature changes a little, i.e., approximate null model, which is also the reason why the augmentation effect of these two models is not always as good as the leaf nodes. Also due to this reason, it has certain stochasticity, e.g., in BZR_MD, ADA-BC has a better effect than the rest of the augmentation models. It is easy to think that this is because the augmentation effect of this dataset is better in this augmentation. Because the importance proportion of BC in COX2 dataset is very high, it can also achieve a weak augmentation effect compared with the zero-order null model although it is an approximate null model.

**Conclusion**

In this paper, we propose a specific data augmentation framework based on the null model to effectively improve the accuracy of graph classification. We conduct experiments to prove the effectiveness of our framework and analyze the experimental results in detail. The experimental results show that the application of data augmentation based on the null model can effectively improve the accuracy of graph classification, and the specific feature data augmentation based on the null model has a better effect than the general null model augmentation. Nowadays, there are numerous researches on data augmentation based on graph classification, but few focus on the important features of graph classification. In the future, we aim to explore the important features of graph data in graph classification in order to achieve more efficient data augmentation.

**References**

2012. Subgraph Matching Kernels for Attributed Graphs. Eprint Arxiv.

Bavelas, A. 1950. Communication patterns in task-oriented groups. The Journal of the Acoustical Society of America, 22(6): 725–730.

Beauchamp, M. A. 1965. An improved index of centrality. Behavioral Science, 10(2): 161–163.

Biddick, M.; and Burns, K. C. 2021. A simple null model predicts the island rule. Ecology Letters.

Bonacich, P. 1972. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology, 2(1): 113–120.

Bonacich, P. 2007. Some unique properties of eigenvector centrality. Social Networks, 29(4): 555–564.

Borgatti, S. P.; and Everett, M. G. 1997. Network analysis of 2-mode data. Social Networks, 19(3): 243–269.

Breiman, L. 2001. Random forests. Machine Learning, 45(1): 5–32.

Carrer, M.; Castagneri, D.; Popa, I.; Pividori, M.; and Lingua, E. 2018. Tree spatial patterns and stand attributes in temperate forests: The importance of plot size, sampling design, and null model. Forest Ecology and Management, 407: 125–134.

Cazabet, R.; Borgnat, P.; and Jensen, P. 2017. Enhancing space-aware community detection using degree constrained spatial null model. In International Workshop on Complex Networks, 47–55. Springer.

Chung, F.; and Lu, L. 2002a. The average distances in random graphs with given expected degrees. Proceedings of the National Academy of Sciences, 99(25): 15879–15882.

Chung, F.; and Lu, L. 2002b. Connected components in random graphs with given expected degree sequences. Annals of combinatorics, 6(2): 125–145.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 113–123.

de Lara, N.; and Pineau, E. 2018. A simple baseline algorithm for graph classification. arXiv preprint arXiv:1810.09155.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Dong, H.; Ding, Z.; He, X.; Feng, F.; and Bi, S. 2020. Data Augmentation View on Graph Convolutional Network and the Proposal of Monte Carlo Graph Learning. arXiv preprint arXiv:2006.13090.

Estrada, E.; Meloni, S.; Sheerin, M.; and Moreno, Y. 2016. Epidemic spreading in random rectangular networks. Physical review E, 94(5): 052316.

Fadaee, M.; Bisazza, A.; and Monz, C. 2017. Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440.
Gjoka, M.; Kurant, M.; and Markopoulou, A. 2013. 2.5 k-graphs: from sampling to generation. In 2013 Proceedings IEEE INFOCOM, 1968–1976. IEEE.

Gotelli, N. J. 2001. Research frontiers in null model analysis. Global ecology and biogeography, 10(4): 337–343.

Gotelli, N. J.; and Graves, G. R. 1996. Null models in ecology.

Jeffrey; J.; Sutherland; Lee; A.; O’Brien; Donald; F.; and Weaver. 2004. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure—Activity Relationships. ChemInform, 35(5).

Lind, P. G.; Gonzalez, M. C.; and Herrmann, H. J. 2005. Cycles and clustering in bipartite networks. Physical Review E, 72(5): 056127.

Mahadevan, P.; Hubble, C.; Krioukov, D.; Huffaker, B.; and Vahdat, A. 2007. Orbis: rescaling degree correlations to generate annotated internet topologies. In Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications, 325–336.

Mahadevan, P.; Krioukov, D.; Fall, K.; and Vahdat, A. 2006. Systematic topology analysis and generation using degree correlations. ACM SIGCOMM Computer Communication Review, 36(4): 135–146.

Menze, B. H.; Kelm, B. M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; and Hamprecht, F. A. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10(1): 1–16.

Narayanan, A.; Chandramohan, M.; Venkatesan, R.; Chen, L.; Liu, Y.; and Jaiswal, S. 2017. graph2vec: Learning distributed representations of graphs. arXiv preprint arXiv:1707.05005.

Newman, M. E. 2006. Finding community structure in networks using the eigenvectors of matrices. Physical review E, 74(3): 036104.

Newman, M. E.; Strogatz, S. H.; and Watts, D. J. 2001. Random graphs with arbitrary degree distributions and their applications. Physical review E, 64(2): 026118.

Nian, X.; and Fu, H. 2014. Efficient routing on two layer degree-coupled networks. Physica A: Statistical Mechanics and its Applications, 410: 421–427.

Pastor-Satorras, R.; Vázquez, A.; and Vespignani, A. 2001. Dynamical and correlation properties of the internet. Physical review letters, 87(25): 258701.

Reimann, M. W.; Gevaert, M.; Shi, Y.; Lu, H.; Markram, H.; and Muller, E. 2019. A null model of the mouse whole-neocortex micro-connectome. Nature communications, 10(1): 1–16.

Ren, Z.-M.; Mariani, M. S.; Zhang, Y.-C.; and Medo, M. 2017. A time-respecting null model to explore the structure of growing networks. arXiv preprint arXiv:1703.07656.

Rokach, L.; and Maimon, O. 2005. Top-down induction of decision trees classifiers - a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 35(4): 476–487.

Şahin, G. G.; and Steedman, M. 2019. Data augmentation via dependency tree morphing for low-resource languages. arXiv preprint arXiv:1903.09460.

Tsitsulin, A.; Mottin, D.; Karras, P.; Bronstein, A.; and Müller, E. 2018. Netlsd: hearing the shape of a graph. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2347–2356.

Tu, K.; Li, J.; Towsley, D.; Braines, D.; and Turner, L. D. 2019. gl2vec: Learning feature representation using graphlets for directed networks. In Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, 216–221.

Ulrich, W.; and Gotelli, N. J. 2013. Pattern detection in null model analysis. Oikos, 122(1): 2–18.

Wolfe, A. W. 1995. Social Network Analysis: Methods and Applications. Contemporary Sociology, 91(435): 219–220.

Xuan, Q.; Wang, J.; Zhao, M.; Yuan, J.; Fu, C.; Ruan, Z.; and Chen, G. 2021. Subgraph Networks With Application to Structural Feature Space Expansion. IEEE Transactions on Knowledge and Data Engineering, 33(6): 2776–2789.

Zhai, X.; Zhou, W.; Fei, G.; Liu, W.; Xu, Z.; Jiao, C.; Lu, C.; and Hu, G. 2018. Null model and community structure in multiplex networks. Scientific reports, 8(1): 1–13.

Zhao, T.; Liu, Y.; Neves, L.; Woodford, O.; Jiang, M.; and Shah, N. 2020. Data augmentation for graph neural networks. arXiv preprint arXiv:2006.06830.

Zhou, J.; Shen, J.; Yu, S.; Chen, G.; and Xuan, Q. 2020. M-Evolve: Structural-Mapping-Based Data Augmentation for Graph Classification. IEEE Transactions on Network Science and Engineering, 8(1): 190–200.