Laboratory (Spring 2023)
HW3- Data Analysis and Prediction on Academic Citation Networks

submitted by: Ofri Hefetz 209028067, Shai Shani 206165318

Git repository link:
https://github.com/OfriHefetz/Data-Analysis-and-Prediction-on-Academic-Citation-Networks

Abstract:
The exponential growth of academic article publications has led to the need for automatic tools that can efficiently categorize articles and make them easily accessible to researchers. This report focuses on predicting the categories of academic articles using machine learning algorithms. A model is developed to automatically classify articles into relevant categories by leveraging a dataset representing a network of academic citations. Exploratory data analysis provides insights into the dataset's composition, patterns, and anomalies, laying the foundation for subsequent model development and optimization. Graph neural network models are evaluated for article categorization, including GCN, GraphSAGE, and GAT. The results highlight the best-performing model as GAT with hidden_channels=64 and num_heads=8, which achieved the highest accuracy among tested configurations. Suggestions for future improvements include incorporating additional evaluation metrics, addressing the class imbalance, and integrating textual and semantic information into the models.

Keywords: academic article categorization, machine learning algorithms, graph neural networks, exploratory data analysis, model selection, optimization.

Introduction:
The exponential growth of academic article publications has presented a challenging problem for researchers to sort through and access extensive scholarly content efficiently. As a result, there is an increasing need for automatic tools that can effectively categorize articles and make them easily accessible to researchers. Machine learning algorithms and statistical tools have emerged as powerful techniques for addressing this challenge. This report aims to predict the categories of academic articles using machine learning algorithms. By leveraging the provided data set, which represents a network of academic citations, we aim to develop a model that automatically classifies articles into relevant categories. The data set consists of a directed graph where each vertex represents an article, and the directed edges between vertices indicate citation relationships. Additionally, each article is associated with a feature vector derived from the abstract and title, representing the article's content. The task involves analyzing, investigating, and classifying the articles based on their category. The main performance metric for evaluating our model will be accuracy, which measures the percentage of correctly predicted categories. We will use a validation set to assess the model's performance and aim for a minimum accuracy requirement of 0.5. By successfully predicting article categories, we can contribute to developing automatic tools that facilitate researchers' access to relevant literature. This report will outline the steps to achieve this goal, including exploratory data analysis, model selection, and optimization.

Exploratory Data Analysis:
Exploratory data analysis was conducted to gain a deeper understanding of the data and uncover valuable insights. This analysis involved the creation of visualizations to enhance comprehension and identify data distribution patterns and anomalies. The following sections summarize the visualizations and key observations during the EDA.

1. Dataset Information:
   the dataset used for this analysis consists of articles with corresponding categories. The following information provides an overview of the dataset:
   - **Number of nodes:** The total number of articles in the dataset -100000
   - **Number of samples in the training se**t: 80000
   - **Number of samples in the validation set**: 20000
   - **Number of features per node:** The number of features extracted from each article, such as abstract and title- 128.
   - **Number of edges:** The number of connections between articles- 444288.
   - **Number of classes:** 40.
   - **Contains isolated nodes:** Indicates whether there are articles that do not have any connections to other articles- True.
   - **Contains self-loops:** - False.
   - **Is undirected**: Indicates whether the connections between articles are bidirectiona - False.

2. Patterns and Anomalies:
   During the exploratory data analysis, several patterns and anomalies in the data distribution were identified. These findings can be crucial for understanding the dataset and improving the accuracy and performance of the classification model. Some of the key observations include:
   - **Category Distribution:** It was observed that certain categories were overrepresented or underrepresented compared to others. This variation in category distribution could potentially impact the model's ability to accurately classify articles.

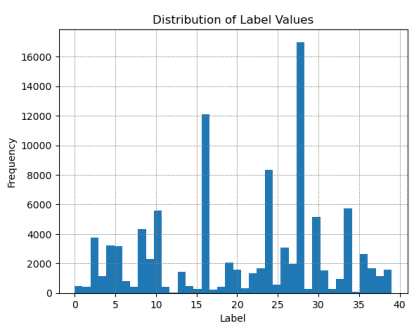

Figure 1 - Labels distributions of all the data

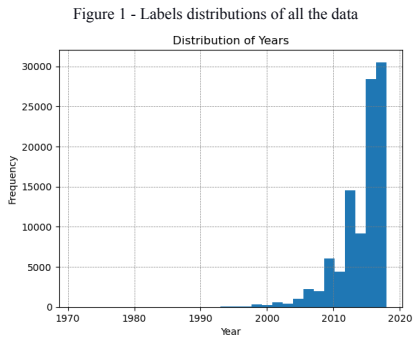

Figure 2 -Years distribution

3. In-Degree and Out-Degree Analysis:
   The in-degree and out-degree of each article were calculated to measure the number of incoming and outgoing connections, respectively. This analysis helps in understanding the popularity and influence of articles within the network. Key observations include:
   - **Average In-Degree:** The average number of incoming connections to an article was found to be 4.44288.
   - **Median In-Degree:** The median number of incoming connections to an article was found to be 1.
   - **Average Out-Degree:** The average number of out connections from an article was found to be 4.44288.
   - **Median Out-Degree:** The median number of out connections from an article was found to be 2.

   By conducting exploratory data analysis and examining various visualizations, valuable insights were obtained regarding the data distribution, patterns, and anomalies. These observations provide a foundation for subsequent model development and optimization steps, allowing for informed decision-making and improved accuracy in predicting article categories using machine learning algorithms.

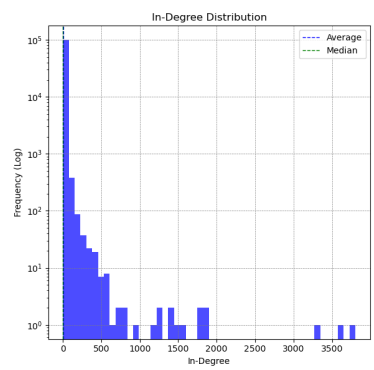

Figure 3 - In degree distribution of all the data

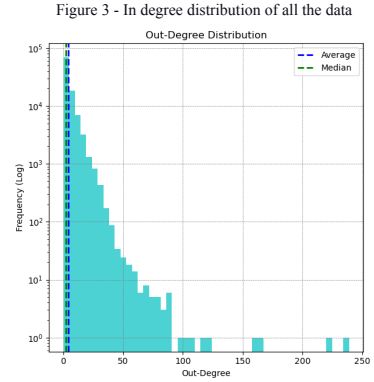

Figure 4 - Out degree distribution of all the data

Example Node Details:

In this section, we will delve into the significance of the features, their representation, and the label's relevance in the context of academic citations. Furthermore, we will elaborate on the in-degree and out-degree metrics and their potential implications.

1. Features and their Representation: The features associated with each node, derived from the abstract and title of academic articles, play a crucial role in capturing the content and characteristics of the articles. These features are typically represented in a numerical or vectorized form that allows for computational analysis. Representing features in academic citations enables extracting meaningful patterns, themes, and topics in the articles. For instance, the feature vector can capture the frequency of certain words or phrases that indicate the article's subject topic through natural language processing techniques. This representation enables classification by providing a structured format for machine learning algorithms to learn from and make predictions based on the feature patterns.
2. Label Relevance and Article Categories: The label assigned to each node represents the category or field of study to which an academic article belongs. The label is typically assigned based on expert knowledge or manual annotation. It is the ground truth for training and evaluating the machine learning model. The label offers the articles a semantic context, allowing researchers to filter and retrieve the most relevant content in their fields of study.
3. Metrics at the In-Degree and Out-Degree Levels: In academic citations, the in-degree and out-degree metrics give useful insights about citation linkages and the influence of works within the network. The in-degree metric refers to the number of incoming citations received by a particular article, indicating the extent to which other articles reference it. Contrarily, the out-degree metric represents the number of outgoing citations from an article, reflecting its references to other works. The in-degree metric can be interpreted as a measure of the popularity and influence of an article within the academic community. Articles with a high in-degree value have been referenced and are likely significant within their field. On the other hand, A higher out-degree value may indicate that an article contributes to expanding knowledge in a particular area.

Model Selection and Testing:

In this section, we will discuss the motivation behind choosing specific models, GCN (Graph Convolutional Networks), GraphSAGE (Graph Sample and Aggregated), and GAT (Graph Attention Networks), and how they are for the task of article categorization. We will also examine different activation functions and their impact on the model's performance.

1. Model Selection: We used the selection of **GCN**, **GraphSAGE**, and **GAT** as the models for article categorization because of their ability to capture and leverage the structural information within the network. These models are designed to handle graph-structured data, making them well-suited for analyzing node relationships and dependencies. **GCN** is a popular choice for graph-based learning tasks due to its simplicity and effectiveness in aggregating information from neighboring nodes. **GraphSAGE**, on the other hand, excels in scenarios with limited labeled data available. It leverages a sampling strategy to aggregate information from a node's local neighborhood, allowing it to learn meaningful representations even in incomplete data. **GAT** introduces attention mechanisms to graph neural networks, enabling it to selectively attend to different parts of the neighborhood when aggregating information. By assigning attention weights to neighboring nodes, **GAT** can focus on the most relevant nodes and effectively capture the importance of different relationships within the citation network.
2. Activation Functions: Different activation functions, such as ReLU (Rectified Linear Unit) and Sigmoid, are applied to the outputs of nodes in the graph neural networks. The choice of activation function impacts the model's ability to capture complex patterns and non-linear relationships within the data.

   ReLU is commonly used in graph neural networks due to its simplicity and ability to mitigate the vanishing gradient problem. Conversely, Sigmoid is often used in binary classification tasks, where it maps the output to a probability value between 0 and 1. In the context of article categorization, ReLU activation can effectively capture non-linear relationships between articles and their features. However, a softmax activation function at the output layer may be more appropriate if the task

involves multi-class classification or predicting probabilities.

3. Hidden Channels and Dropout Values: The hidden_channels parameter determines the dimensionality of the hidden representation at each layer of the graph neural network. Increasing the hidden_channels may enable the model to capture more complex patterns and increase the computational complexity and the risk of overfitting. *Dropout* is a regularization technique used to prevent overfitting in machine learning models. Adjusting the dropout value allows for controlling the regularization applied to the model, balancing model complexity and generalization performance.
4. Learning rate: During the model testing phase, various models were evaluated using different learning rates ranging from 0.0001 to 0.001. The learning rate is a crucial hyperparameter that determines the step size at which the model updates its parameters during training. By exploring a range of learning rates, it was possible to assess their impact on the model's performance.

Results:

This section presents a comprehensive summary of the results obtained from each model configuration, including the accuracy and loss values. We also discuss the best-performing model(s) and their corresponding hyperparameters.

Model Performance Summary: We trained and evaluated various model configurations to classify articles within the academic citation network. Here, we provide a summary of the accuracy and loss values obtained from each model configuration:

Model and parameters	Results
GCN (hidden_channels=30, dropout=0.5,activation=torch.tanh)	Validation Accuracy: 0.5024 Minimum Loss Value: 2.0617
GCN (hidden_channels=80, dropout=0.6, activation=torch.relu)	Validation Accuracy: 0.5655 Minimum Loss Value: 1.6046
GCN (hidden_channels=120, dropout=0.6, activation=torch.relu)	Validation Accuracy: 0.5792 Minimum Loss Value: 1.5202
GCN (hidden_channels=180, dropout=0.6, activation=torch.relu)	Validation Accuracy: 0.5819 Minimum Loss Value: 1.4764
GCN (hidden_channels=200, dropout=0.6, activation=torch.tanh)	Validation Accuracy: 0.5683 Minimum Loss Value: 1.4782
GraphSAGE (hidden_channels=64, activation=torch.relu)	Validation Accuracy: 0.5612 Minimum Loss Value: 1.7085
GraphSAGE (hidden_channels=180,activation=torch.relu)	Validation Accuracy: 0.5777 Minimum Loss Value: 1.5353
GAT (hidden_channels=64, num_heads=8,activation=torch.relu)	Validation Accuracy: 0.5924 Minimum Loss Value: 1.3829

The above results highlight the performance of each model configuration in terms of accuracy and loss. Different hyperparameters and model architectures produce varying levels of performance.

Best-Performing Model(s) and Hyperparameters: Based on the results obtained, the best-performing model(s) for classifying articles within the academic citation network are: GAT (Graph Attention Network) with hidden_channels=64 and num_heads=8, relu activation function and dropout rate of =0.6:

- Validation Accuracy: 0.5924
- Minimum Loss Value: 1.3829

This model configuration achieved the highest accuracy among all the tested configurations. The selection of hidden_channels and num_heads parameters was crucial in achieving improved performance. In conclusion, based on our testing and evaluation results, we have selected the GAT algorithm as the preferred model for article categorization within the academic citation network. Its ability to provide better accuracy and effectively capture the intricate relationships between articles outweighs the computational cost. Adopting the GAT algorithm will contribute to developing automatic tools that enhance researchers' access to relevant literature and aid in knowledge discovery and exploration.

Conclusion:

In conclusion, our analysis of various model configurations for automatic article categorization within the academic citation network produced several key findings and insights. The best-performing model was identified as the GAT (Graph Attention Network) with hidden_channels=64 and num_heads=8, which achieved the highest accuracy among the tested configurations.

However, it is vital to acknowledge the difficulties and limitations of our method. Our examination was limited to the performance of a few different model parameters, with insufficient weight given to other factors like class imbalance or any biases in the data. Addressing these factors and conducting further investigations would strengthen the validity and robustness of the automatic article categorization process.

For future improvements or research directions, we suggest the following:

1. Incorporate additional evaluation metrics: Alongside accuracy and loss values, utilize precision, recall, and F1-score to evaluate the models' performance across different categories comprehensively.
2. Address class imbalance: Investigate techniques to handle class imbalance within the dataset, such as oversampling minority classes or employing class-weighted loss functions. This can help improve the performance of the models in categorizing articles from underrepresented categories.
3. Incorporate textual and semantic information: Integrate text-based features and semantic embeddings into the models to leverage the content of the articles and capture deeper contextual information, thus improving the accuracy of article categorization.

In conclusion, while our analysis identified a high-performing model and shed light on the automatic article categorization process, further research and improvements are needed to overcome the limitations and enhance the effectiveness of the models in this field.

Sources of information:

[1]Hands-On Guide to PyTorch Geometric
[2]pytorch_geometric documentation
[3] Graph Attention Networks in Python | Towards Data Science
[4] CS224W: Machine Learning with Graphs | 2021 | Lecture 6.1 - Introduction to Graph Neural Networks