Optimizing Dataset Quality for Enhanced Machine Learning Performance

The Impact of Dataset Metrics

AUTHOR

Efe Unluyurt

RESPONSIBLE PROFESSOR

Kubilay Atasu

SUPERVISOR

Atahan Akyildiz

TU Delft

INTRODUCTION

The quality and characteristics of datasets play a pivotal role in the effectiveness of machine learning models, especially where GNNs, LLMs, and Transformers are used.

This project explores the important aspects that make a dataset valuable and proposes ideas about understanding such datasets. Understanding and improving dataset quality is crucial because the data's integrity directly influences the performance and reliability of the models.

RESEARCH QUESTIONS

- What makes a good dataset?
- Which metrics best describe a dataset's usefulness?

BACKGROUND INFORMATION

Previous research such as Open Graph Benchmark: Datasets for Machine Learning on Graphs'' has explored various aspects of dataset construction, evaluation, and augmentation. This research aims to fill the gap by conducting an in-depth exploration of different datasets and their applicability to various machine learning tasks, ultimately finding out the dataset metrics that make a great dataset for a specific task

Metrics such as the percentage of missing cells and graph sparsity are important to evaluating dataset quality. These metrics could help in understanding the potential impact of data imperfections on machine learning models.

RELATED LITERATURE

Key references related to the research include:

- Wei Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hengrui Luo, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. 2020.
- Imran Razzak Adnan Ejaz and Saeed Anwar. A survey on dataset quality in machine learning. Engineering Applications of Artificial Intelligence, 2023.

METHODOLOGY

The study involves:

- Retrieving various datasets
- Preprocessing them to ensure suitability for ML tasks
- Choosing metrics that will be evaluated
- Selecting the most appropriate datasets/models to evaluate those metrics
- Conducting different experiments to evaluate these metrics

The datasets explored include OGB-Arxiv, Amazon-Review, Ethereum phising transaction network, and IBM-AML datasets.

Dataset	Type	Data Content
Amazon-Fashion	Text-based, graph structured	Amazon customer reviews and ratings on fashion products
IBM-AML	Numeric, graph structured	Synthetic transaction records with flags indicating suspicious transactions.
Ogbn-arxiv	Numeric, graph structured	A citation network of Computer Science (CS) papers from arXiv. (Text values are represented as node2vec vectors)
Ogbn-arxiv (text)	Text-based, graph structured	A citation network of Computer Science (CS) papers from arXiv.
Ethereum Phishing Transaction Network	Numeric, graph structured	Transaction records related to Ethereum phishing activities

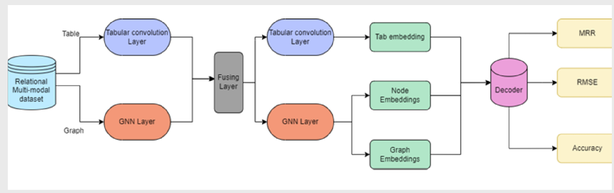
ANALYSIS/EXPERIMENTS

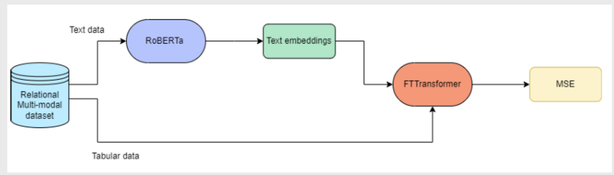
Four metrics were chosen to be analysed; Graph Sparsity, Percentage of Empty Cells, Modularity, Text Length

The datasets and models were chosen to evaluate the effect of these metrics on the model's performance.

IBM-AML -> Graph Sparsity, Empty Cells, Modularity
Amazon Fashion -> Text Length

The datasets where then modified and different runs were taken to evaluate the metrics and to see if they affect the outcome of the model.





RESULTS/FINDINGS

Graph Sparsity:

- Too sparse graphs kept information better, resulting in higher accuracy (0.838) and lower RMSE (0.110).
- High density did not improve performance due to added noise and complexity.

Number of Missing Cells:

- The model showed resilience to missing data, keeping high accuracy and low RMSE even with up to 50% missing cells.
- This might be because the model could rely more on the patterns, or the imputation techniques could help to keep the resilience.

Modularity:

- Higher modularity generally improved model performance by resulting in great link prediction.
- However, excessive modularity introduced complexity and decreased accuracy (0.813) and increased RMSE (0.116) of the model.

Text Length:

- Longer texts (>= 70 words) provided more and quality context, leading to the lowest MSE (0.3502), but it is still great to keep a balance since the noise in text data can create challenges.
- Shorter texts (<= 10 words) led to higher MSE (0.3663), indicating the importance of rich textual information for model accuracy.

CONCLUSION/FUTURE IMPROVEMENTS

The research highlights the important role of dataset quality in enhancing machine learning model performance, and analyses some metrics to find out to what extent they effect the dataset quality. The findings include:

- **Graph Sparsity:** Maintaining an optimal level of sparsity preserves essential relationships and improves model accuracy.
- **Missing Data:** Models can effectively handle missing data with proper imputation techniques, maintaining high performance even with significant data gaps.
- **Modularity:** While useful for identifying patterns, excessive modularity can introduce complexity that hinders performance.
- **Text Length:** Providing adequate textual context is essential for the accuracy of language models.

Future research can go deeper into advanced data imputation methods for empty cells, finding optimal modularity levels, and conduct the experiments for other metrics such as graph diameters. Apart from those, the further studies can be made to evaluate the results provided by this study, like why and how does modularity improves model performance.