Improving the Generalisation Performance of the Weisfeiler-Lehman Subtree Kernel

Data Management Plan

Lead partner	RWTH Aachen – Christopher Morris
Version	1.0
Status	Done
Dissemination level	Public
Project DOI	10.5281/zenodo.11081253

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
1.0	28.02.2024	▪ Final version

Abstract

This document is a data management plan (DMP) for the bachelor thesis project “Improving the Generalisation Performance of the Weisfeiler-Lehman Subtree Kernel”, done in Q1 and Q2 of 2023 at RWTH Aachen, written by Antoine Origer ORCID ID 0009-0006-7102-8571 and tutored by Christopher ORCID ID 0000-0002-0465-1068.

This DMP addresses all important aspects around the data used and created in this thesis.

Content

1. Data Summary

Will you re-use any existing data and what will you re-use it for?

For this project, we are reusing several, well known, existing datasets, including AIDS, PROTEINS, MUTAG, Mutagenicity, IMDB-BINARY, SYNTHETICnew, COIL-DEL, and NCI1. These datasets are integral to the research as they are open to free use and they provide a foundation for assessing the generalization performance of the Weisfeiler-Lehman subtree kernel in chemoinformatics and machine learning contexts. The reuse of these datasets ensures robust testing across diverse data types, enhancing the project's reliability and applicability.

Dataset name	Amount of Graphs	Type of graphs	File format	Approximated volume	Containing sensitive information
AIDS	2000	Small molecules	.txt	3 MB	No
PROTEINS	1113	Bioinformatics	.txt	3 MB	No
MUTAG	188	Small molecules	.txt	150 KB	No
Mutagenicity	4337	Small molecules	.txt	5 MB	No
IMDB-BINARY	1000	Social networks	.txt	5 MB	No
SYNTHETICnew	300	Synthetic	.txt	2 MB	No
COIL-DEL	3900	Computer vision	.txt	7 MB	No
NCI1	4110	Small molecules	.txt	4 MB	No

What types and formats of data will the project generate or re-use?

The project will predominantly generate and utilize data in text file format (TXT) for graph representations and Excel files (XLSX) for accuracy measures. These formats are chosen for their simplicity and compatibility with existing tools used in the research.

As for the coding part of this project, all source code will be written in Python, as is it well suited and convenient for this type of work.

File(s) created	Type	File format	Approximated (total) volume	Containing sensitive information
programming files (22)	Source code	Python .py	< 1 MB	No
Bachelor thesis	Paper	PDF	550 KB	No
Result data	Accuracies of SVM models	.xslx	150 KB	No

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The purpose of data generation and re-use directly supports the project's objective to enhance the Weisfeiler-Lehman subtree kernel's generalization performance. Data in for of dataset is needed to run, test and evaluated the quality of a model. Gathering the resulting accuracies is done for comparation, pattern finding and generally improving the understanding of the model. By using and generating data, the project aims to comprehensively evaluate and optimize the kernel's efficiency and accuracy across varied datasets.

What is the origin/provenance of the data, either generated or re-used?

The datasets are well-known in the research community, sourced from public databases and previously utilized in related chemoinformatics research. All the mentioned datasets in the first table were provided by the tutor of this thesis, Christopher Morris, on graphlearning.io

Provenance is further documented through the README file containing source references.

To whom might your data be useful ('data utility'), outside your project?

The data and findings could be valuable to researchers in machine learning and chemoinformatics, educators in computational chemistry, and developers of similar computational models seeking to validate or enhance their algorithms.

2.FAIR data

2.1 Making data findable, including provisions for metadata

A DOI will be assigned to the project's dataset and code repository upon completion, ensuring a persistent identifier is available. Metadata will include detailed descriptions of the dataset, methodology, and computational requirements housed within a README file. No specific metadata standards are applicable to this discipline; however, common metadata elements like title, description, and keywords will be incorporated following general academic practices.

2.2 Making data accessible

The datasets will be openly available on GitHub, which also serves as a platform for version control and collaboration, ensuring the data is both accessible and identifiable. GitHub's infrastructure allows the data to be assigned a unique DOI, and the platform itself supports open access protocols. GitHub is considered an excellent platform to share code or projects publicly or in restricted groups. It is today the largest code host in the world, and we consider it perfectly suited for making this project openly available. As for the programming language, Python 3 are extremely widely known and easy to setup with an IDE of choice. The necessary extensions can be found in the source code.

2.3 Making data interoperable

Data and metadata will be maintained in widely used formats such as TXT for graph data and XLSX for accuracy metrics, ensuring compatibility across various systems. We aim to adhere to community-endorsed best practices by maintaining clear and comprehensive documentation as well as controlled vocabulary and naming conventions, enabling data exchange and reuse.

2.4 Increase data re-use

All data will be published under an open license (CC-BY 4.0 for data, MIT for software), maximizing re-use potential. Documentation will include all necessary details to replicate the study and validate findings, such as methodologies, codebooks, and variable definitions.

3. Allocation of Resources

All resources required, such as personal computing power and online data storage through GitHub and Google Drive, are already available and managed by the project's sole researcher, Antoine Origer.

No additional financial resources are available or necessary for the project or its management of data.

4. Data Security

Data security measures include routine backups to both Google Drive and GitHub. As the data does not contain sensitive information, standard security protocols on these platforms are adequate. As this project aims to publish everything and make all the data freely accessible, no further security measures are deemed necessary. All these data management tasks will be done by Antoine Origer using his own necessary accounts.

5. Data preservation

All Data being uploaded on GitHub assures that it should stay preserved for at least 10 years. There are no plans to remove the project from GitHub at any point.

6. Ethics and legal

The project does not involve any personal or sensitive data, thereby minimizing ethical concerns. The open data and software licenses (CC-BY 4.0 and MIT) ensure proper citation and reuse in line with ethical academic standards.