DATA AUGMENTATION IN GRAPH NEURAL NETWORKS: THE ROLE OF GENERATED SYNTHETIC GRAPHS

Sumeyye Bas, **Kiymet Kaya,** Resul Tugay, Sule Gunduz Oguducu
bass20@itu.edu.tr, **kayak16@itu.edu.tr**, resultugay@gazi.edu.tr, sgunduz@itu.edu.tr

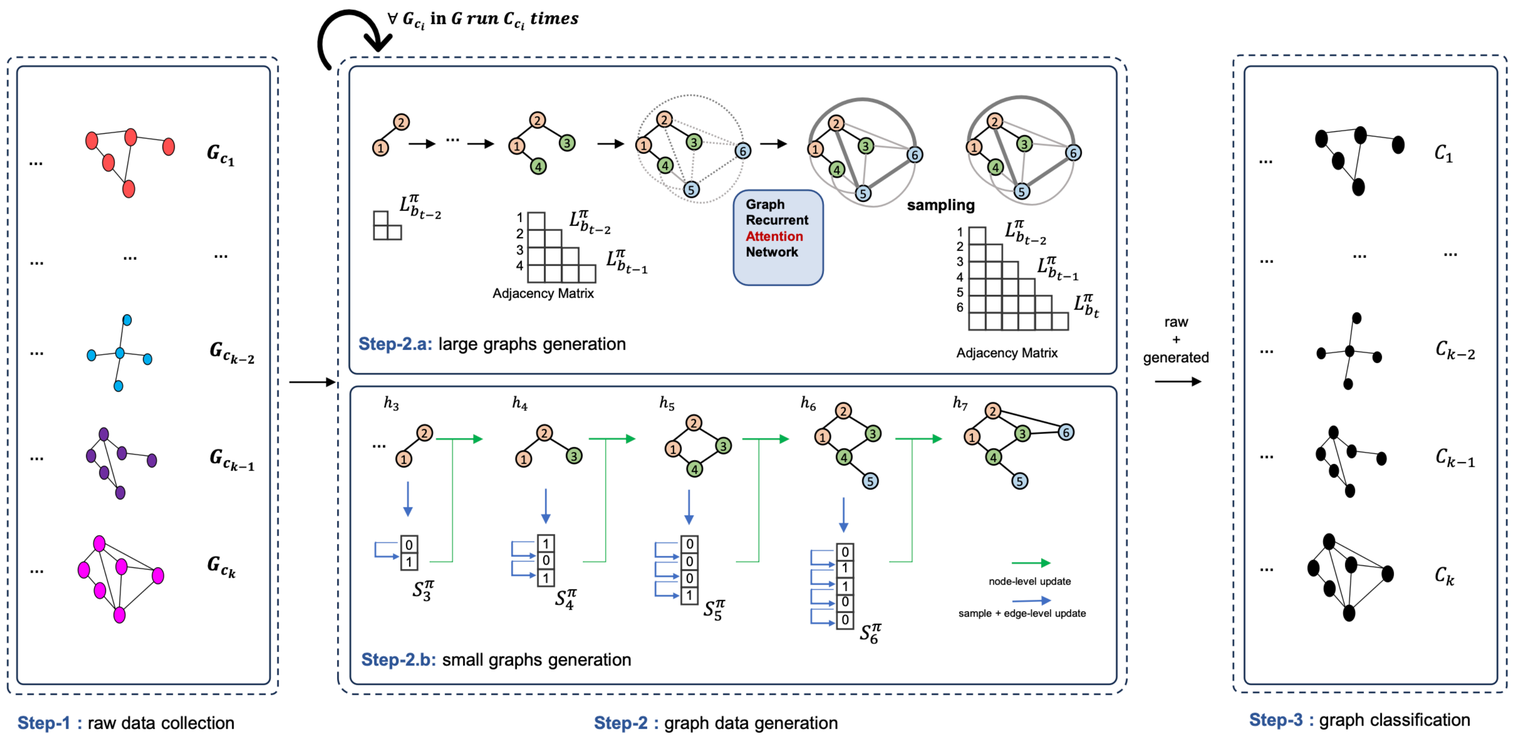
Introduction

Graphs are important representations for interrelated data in social networks and bioinformatics. These complex structures capture deep relationships with ease, making predictive modeling possible. Thus, attaining high-quality graph representation is critical for identifying underlying interrelated patterns and events in domains involving graph-related data. However, the progress in this area is being hampered by the lack of useful big datasets and consistent evaluation processes. AI-generated data is becoming more and more popular for a variety of uses

Contributions

- Examining the usability of Al-generated data to improve Al model performances.
- Investigating how the ratio of Al-generated data affects the prediction capabilities of the models by training the models with different proportions of synthetic data.
- Addressing the labeling issue in graph data augmentation.
- Providing a framework of label invariant graph data augmentation for improving graph classification performance.

Proposed Framework



1. Graphs in the input dataset are separated by graph labels.
2. Then, corresponding generator models are trained for each class.
3. Both original and generated graphs are used in training graph classification.

Experiments

- We observed the effect of AI-generated graph data by GNNs on graph classification on six public dataset: Three chemical compounds (MUTAGENICITY, ENZYMES, MUTAG), two social networks (COLLAB, TWITCH EGOS), and one protein interactions (DD).
- We partitioned each dataset into three (80-10-10%), ensuring the class distributions remained consistent:
  - Raw-data: Initial data available to researchers.
  - Subreal data (R): Imitates the supplementary data that researchers might acquire through additional time and resource investment in real-world scenarios.
  - Test data: is designated to evaluate the performance of graph classification.
- We trained separated generators for each class in training set and assigned labels corresponding to the generators.

Table1

		GraphSAGE		GIN0		GINWithJK		GCNWithJK		EdgePool	
		Acc.	Epoch	Acc.	Epoch	Acc.	Epoch	Acc.	Epoch	Acc.	Epoch
DD	raw-data	0.630	6	0.605	6	0.545	27	0.630	5	0.660	4
	w/ Real = $\|R\|$	0.630	12	**0.665**	40	0.630	12	0.630	4	0.630	5
	w/ Gen.$_1$ = $\|R\|$	0.635	10	0.650	14	0.665	20	0.630	4	0.630	5
	w/ Gen.$_2$ = $\|R\| * 2$	**0.695**	31	**0.665**	13	0.665	23	0.630	4	0.630	4
	w/ Gen.$_3$ = $\|R\| * 3$	0.630	14	**0.665**	15	**0.685**	12	0.630	5	0.630	5
COLLAB	raw-data	0.576	7	0.718	40	0.707	11	0.534	8	0.650	5
	w/ Real	0.572	10	0.736	28	**0.736**	10	**0.693**	13	0.656	16
	w/ Gen.$_1$	**0.578**	8	0.705	9	0.718	11	0.534	6	0.666	14
	w/ Gen.$_2$	0.572	14	0.730	10	0.734	11	0.650	24	**0.701**	19
	w/ Gen.$_3$	0.569	13	**0.738**	33	0.716	8	0.631	13	0.627	23
TWITCH EGOS	raw-data	0.576	16	0.681	20	0.701	39	0.689	30	0.682	45
	w/ Real	0.576	7	0.694	31	0.700	30	0.684	13	OOM	
	w/ Gen.$_1$	0.578	10	**0.700**	24	**0.702**	23	0.695	21	OOM	
	w/ Gen.$_2$	**0.579**	12	0.699	12	**0.702**	28	0.696	19	OOM	
	w/ Gen.$_3$	0.575	9	0.699	13	0.690	42	**0.700**	35	OOM	
MUTAGENICITY	raw-data	0.597	20	0.703	15	**0.740**	28	0.551	4	0.701	28
	w/ Real	0.590	14	0.726	16	0.719	12	0.551	4	**0.719**	44
	w/ Gen.$_1$	0.593	29	0.698	10	0.726	22	0.551	4	0.717	30
	w/ Gen.$_2$	**0.609**	18	0.708	11	0.728	14	0.551	4	0.701	32
	w/ Gen.$_3$	0.586	12	**0.744**	32	0.712	31	0.551	4	0.685	22
ENZYMES	raw-data	0.141	9	0.166	7	0.191	46	0.166	13	0.166	10
	w/ Real	0.166	8	0.158	15	0.166	96	0.166	8	0.166	17
	w/ Gen.$_1$	**0.233**	14	0.200	13	**0.266**	41	0.166	10	**0.175**	15
	w/ Gen.$_2$	0.191	7	**0.233**	27	**0.266**	35	0.166	12	**0.175**	15
	w/ Gen.$_3$	0.175	11	0.208	18	0.158	22	0.166	11	0.166	15
MUTAG	raw-data	0.666	22	0.944	25	0.944	26	0.666	10	0.666	11
	w/ Real	0.666	32	0.833	54	0.833	15	0.666	11	0.666	10
	w/ Gen.$_1$	0.666	12	**0.944**	42	0.833	14	0.666	11	0.666	11
	w/ Gen.$_2$	0.666	11	**0.944**	51	**0.944**	24	0.666	11	0.666	10
	w/ Gen.$_3$	0.666	40	0.833	25	**0.944**	36	0.666	13	0.666	11

Table1 shows the outcomes of 3 techniques differ by the amount of generated graphs. In this experiments, label distribution is saved.

- **w/Gen1:** The same number of generated graphs as the size of R
- **w/Gen2:** 2 times size of the R
- **w/Gen3:** 3 times size of the R

Table2

		GraphSAGE		GIN0		GINWithJK		GCNWithJK		EdgePool	
		Acc.	Epoch	Acc.	Epoch	Acc.	Epoch	Acc.	Epoch	Acc.	Epoch
MUTAGENICITY	raw-data	0.597	20	0.703	15	0.740	28	0.551	4	0.701	28
	w/ Real	0.590	14	**0.726**	16	0.719	12	0.551	4	**0.719**	44
	w/ Gen. (GraphRNN)	**0,611**	20	0,710	14	**0,744**	23	0,551	6	0,694	35
	w/ Gen. (GRAN)	0,551	7	0,689	21	0.728	23	0,551	5	0,551	6
ENZYMES	raw-data	0.141	9	0.166	7	0.191	46	0.166	13	0.166	10
	w/ Real	0.166	8	0.158	15	0.166	96	0.166	8	0.166	17
	w/ Gen. (GraphRNN)	0,166	12	**0,241**	49	**0,241**	65	0,166	14	0,166	15
	w/ Gen. (GRAN)	**0,183**	7	0,158	31	0,125	75	0,166	22	**0,175**	30
MUTAG	raw-data	0.666	22	**0.944**	25	0.944	26	0.666	10	0.666	11
	w/ Real	0.666	32	0.833	54	0.944	15	0.666	11	0.666	10
	w/ Gen. (GraphRNN)	0,388	70	0,888	85	**0,944**	79	0,666	14	**0,777**	52
	w/ Gen. (GRAN)	**0,944**	27	0,666	30	0,777	60	0,666	14	0,777	21

- Table2 shows the outcomes when we drastically increased the proportion of the generated data size for datasets having a few number of small graphs. Generator is specified according to the size of graphs.

Conclusion

To conclude that, our study demonstrates the substantial impact of AI-generated graph data on the performance of graph classification tasks across diverse datasets. The proposed AI-based Generation Framework offers a flexible graph data generation process which is applicable for small, medium, large sized graphs. Moreover, the experiments involving a significant increase in the proportion of generated data further highlighted the potential of our graph generation framework to mitigate issues of data imbalance and scarcity, especially in smaller datasets.