Graph Neural Networks for Missing Data Imputation

Candidate: **Indro Spinelli**

Advisors: Prof. A. Uncini, Dr. S. Scardapane

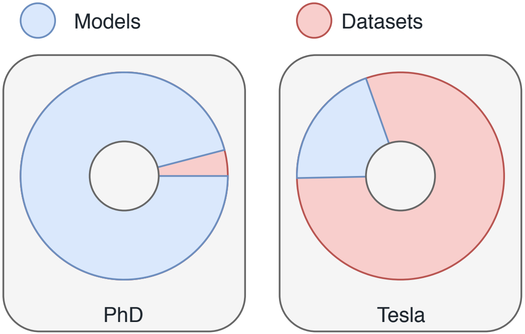
Sapienza University of Rome

Introduction

Dataset importance in ML

The availability of high-quality training **datasets** is a key factor for deploying **machine learning** models in the real world.

Amount of lost sleep by Andrej Karpathy over...



The Problem of Missing Data

Many real-world datasets are affected by the problem of **missing values**.

ID	A	B	C
1	2.7	cat	0
2	0.5	**NaN**	1
3	**NaN**	dog	**NaN**

Fitting a model or performing training with a dataset that has a lot of missing values can drastically impact the model's quality.

For this reason, the field of Missing Data Imputation (MDI) has attracted significant attention.

State Of The Art

A great number of methods have been proposed to solve the MDI task.

- mean imputation Little *et al.* 1986
- MICE Van Buuren *et al.* 2011
- k-nearest neighbors Acuna *et al.* 2004
- random forest Stekhoven *et al.* 2011
- linear models Lakshminarayan *et al.* 1996
- matrix factorization Mnih *et al.* 2008
- support vector machines Wang *et al.* 2006
- neural networks Yoon *et al.* 2018

We will use them to validate the performance of our contribution.

Predictive Imputation

Many of these methods derive from **classical** machine learning algorithms (e.g., for regression and classification) with a few modifications.

This is possible because data imputation can be framed under a predictive framework[1].

Some of these algorithms build a **global** model for data imputation, others instead, use **similar** data points to infer the missing components.

[1] Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. "From Predictive Methods to Missing Data Imputation: An Optimization Approach.". In: *Journal of Machine Learning Research* 18 (2017), pp. 196–1.

Contribution

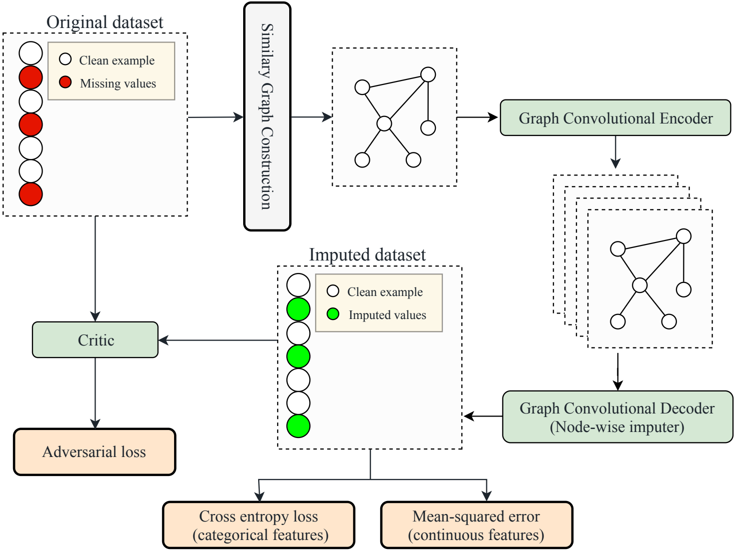
Our contribution

The contribution of this thesis work consists of a new framework for MDI, **GINN**[2](Graph Imputation Neural Network), that exploits both **similar** data points for each imputation and a **global** model built from the overall data set.

This is possible thanks to a new class of **neural network** that is able to model and exploit structured information (in the form of relationships between samples), by working in the domain of **graphs**.

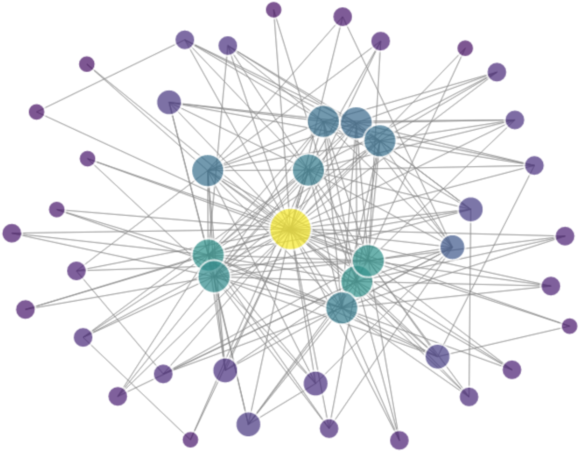
[2] Indro Spinelli, Simone Scardapane, and Aurelio Uncini. "Missing Data Imputation with Adversarially-trained Graph Convolutional Networks". In: *Submitted to Neural Networks (Elsevier)* (2019).

GINN Schematics



An Example on Iris

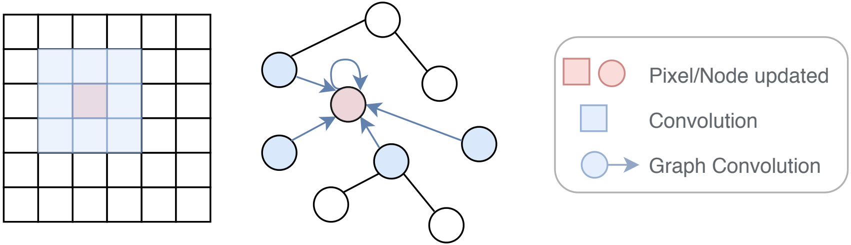
The **similarity graph** describes the structural proximity between samples. For each node the color intensity represents the number of connections and the size the number of missing features.



Graph convolutional layer

The **graph convolutional layer**[3] is the fundamental building block of our graph autoencoder.

2D Convolution vs Graph Convolution



🟥 ● Pixel/Node updated

🟦 Convolution

○→ Graph Convolution

[3] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *Proc. 2017 International Conference on Learning Representations (ICLR)*. 2017.

Graph Autoencoder

Autoencoder

$$\mathbf{H} = \sigma\left(\mathbf{X}\boldsymbol{\Theta}_1\right) ,$$
$$\widehat{\mathbf{X}} = \sigma\left(\mathbf{H}\boldsymbol{\Theta}_2\right) .$$

Graph Autoencoder

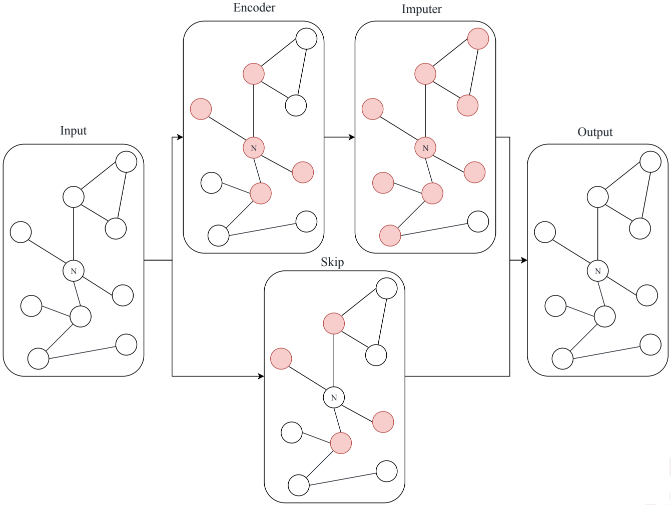
$$\mathbf{H} = \sigma\left(\mathbf{L}\mathbf{X}\boldsymbol{\Theta}_1\right) ,$$
$$\widehat{\mathbf{X}} = \sigma\left(\mathbf{L}\mathbf{H}\boldsymbol{\Theta}_2\right) .$$

- $\mathbf{L}$ is the normalized version of the graph Laplacian.
- The **Laplacian** is a matrix representation of a graph.
- $\mathbf{LX}$ propagates the information across the 1-hop neighbors.
- $\mathbf{LH}$ propagates the information across the 2-hop neighbors.

Our's Graph Convolutional Autoencoder

$$\mathbf{H} = \mathrm{ReLU}\left(\mathbf{L}\mathbf{X}\boldsymbol{\Theta}_1\right) ,$$

$$\widehat{\mathbf{X}} = \mathrm{Sigmoid}\left(\mathbf{L}\mathbf{H}\boldsymbol{\Theta}_2 + \widetilde{\mathbf{L}}\mathbf{X}\boldsymbol{\Theta}_3\right) .$$



Adversarial Training

To improve the quality of the imputed values, we use an adversarial training strategy where a **critic**, a feedforward network in our case, learns to distinguish between **imputed** and **real** data.

The autoencoder is thus trained to **fool the critic** with an additional component in the loss function.

Having an adversarial loss during reconstruction forces the imputed vector to lie close to the natural distribution of the original data.

Losses

Graph Convolutional Autoencoder

$$\mathsf{L}_{Reconstruction} = \alpha \mathbf{MSE}(\mathbf{X}, \widehat{\mathbf{X}}) + (1 - \alpha)\mathbf{CE}(\mathbf{X}, \widehat{\mathbf{X}}) \, ,$$

$$\mathsf{L}_{Total} = L_{Reconstruction} + \lambda L_{Adversarial} \, .$$

- **MSE**$(\mathbf{X}, \widehat{\mathbf{X}})$ is the mean squared error for numerical variables.
- **CE**$(\mathbf{X}, \widehat{\mathbf{X}})$ is the cross entropy loss for categorical variables.

Experimental Evaluation

Metrics

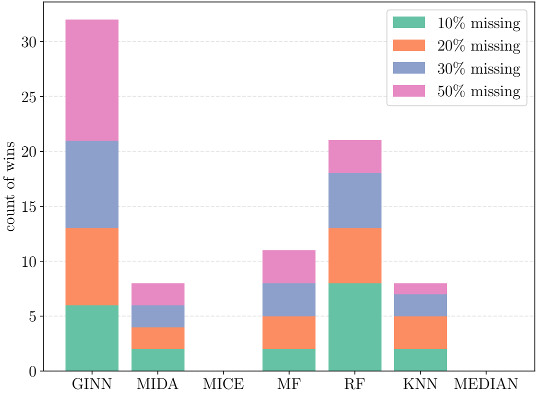
We used a selection of 20 datasets having numerical, categorical and mixed feature vectors, by applying 4 different levels of missing elements: 10%, 20%, 30%, and 50%. We compared **GINN** against 6 state-of-the-art competitors.

First, we evaluated the **reconstruction performances** in terms of RMSE and MAE.

Then we evaluated the **accuracy of post-imputation prediction** with 4 different classifiers in order to see which imputation methods permitted to achieve the best accuracy over a common undamaged test set.

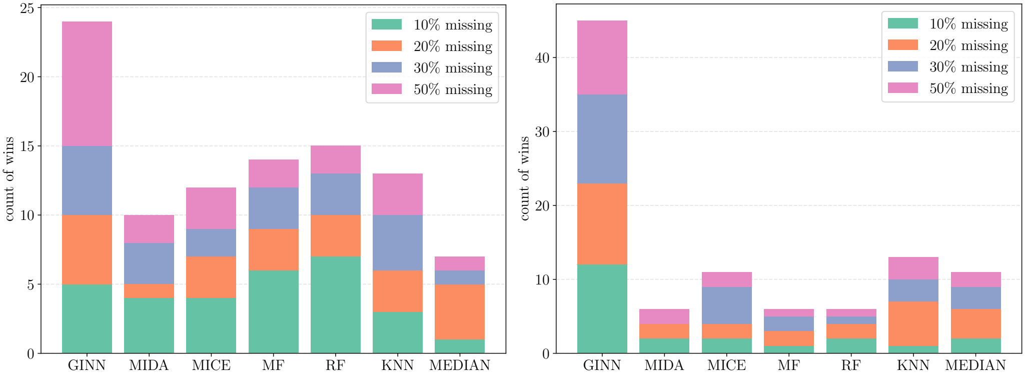
This resulted in a total of $20 \times 4 \times 5 = 400$ experiments, each one ranging from 2 minutes to 12 hours approximately.

Reconstruction Performances



Number of datasets in which each MDI method achieves lowest average RMSE from the true values. The different colors of the bar stand for different percentages of missing elements: from the bottom 10% at the lowest to the top 50% at the highest.

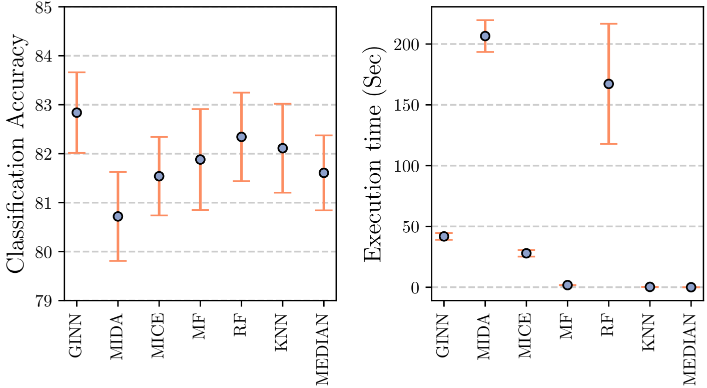
Predictive Performances



Number of datasets in which each MDI method achieves the highest classification accuracy for, k-nn (left) and random forest (right) classifiers.

Real World Missingness

We evaluate the performance of GINN also on real-world datasets with pre-existing (i.e., not artificially induced) missing values. We show the accuracy and computational cost of the method when compared to other state-of-the-art approaches on the Mammographic mass dataset.



Conclusions

Conclusions

- The main contribution of this thesis work is the introduction of a novel technique for missing data imputation, where we used a new graph convolutional autoencoder to reconstruct the full dataset starting from the damaged one.

- We showed through an extensive numerical simulation that our method significantly outperforms state-of-the-art approaches for missing data imputation, especially for large values of noise.

Conclusions

- Different architectures.
- The extension to other types of noisy data beyond tabular data like images and time series.
- Training our imputation module together with a classification step in an end-to-end fashion.

Thank you!