A Comparative Study of Imputation Methods for Time Series Data

Daniyal Khan, Alina Lazar
Youngstown State University
Youngstown, OH, USA
dkhan@student.ysu.edu, alazar@ysu.edu

Abstract

Missing and incomplete values pose a significant challenge in analyzing tabular and time-series data. Dealing with missing values is time-consuming and tedious, especially when working with data from real-world applications. While some imputation approaches estimate missing values based on existing observations, these methods often rely on strong assumptions about the data distribution, which only sometimes improves downstream accuracy. Although tabular imputation methods can be applied to time-series data, incorporating the time component can enhance accuracy. This study evaluates various techniques for missing data imputation in time-series data. We run experiments on four multi-variate time series datasets using five imputation methods. We report training time and testing accuracy.

Introduction and Background

The problem of missing values in datasets is a concern in many domains where tabular and time-series data plays an essential role. Such areas include survey sampling, finance, signal processing, etc. In today's world, where complex interconnected systems like sensor networks or Internet of Things devices prevail, faulty sensors and network failures are common occurrences that disrupt the data acquisition process. Fortunately, these types of failures tend to be sparse and confined to individual sensors, rather than compromising the entire network at once.

Numerous studies have attempted to identify good solutions to recover missing values. In the past, feature imputation as well as label prediction techniques were tackled using popular statistical methods such as interpolation and auto-regressive models. However, these methods have several limitations, including relying heavily on assumptions about the data, being inflexible when handling categorical and continuous data, and lacking generalized approaches that could handle unseen data. Additionally, statistical methods are not capable of leveraging relational information such as the spatial and temporal components of time-series data, which could potentially result in more accurate and dependable predictions in the downstream tasks.

Recently, several deep learning approaches to solve the problem also came into the picture which disregarded temporal information present in the data and implemented simplistic neural network architectures, which were originally developed for sequence processing, to solve the missing value problem in MTSA. Approaches such as Autoencoders or Generative Adversarial Network (GANs), on the other hand, attempts to learn the comprehensive distribution of multivariate time-series data which is eventually used to determine the missing values. While generating flexible solutions they lack in comprehensive application of other observations and use specific initial default values which, eventually, causes biased assumptions. Representing the tabular data and time-series data as a graph makes the problem more understandable and predictable by the application of Graph Neural Networks (GNNs) [6]. GNNs helped to take the temporal as well as spatial components of the data into account while predicting the missing values and also ensures beneficial utilization of neighboring observations.

The choice of the most suitable method depends on factors such as the amount of missing data, the complexity of the data, and the intended downstream analysis.

In this work we are studying missing data imputation techniques and comparing them based on different accuracy measurement metrics. In order to perform realistic experiments we are simulating four types of missing values. These values are generated as binary masks and applied to non-missing datasets. After the values are recovered, the accuracy in terms of mean absolute error (MAE), root mean square error (RMSE), and Mean squared error (MSE) are calculated and reported.

1. **Missing at Random (MAR)**, Missing values are only dependent on observed values
2. **Missing Completely at Random (MCAR)**, the missing values are independent of other values in the data-set
3. **Missing Not at Random (MNAR)**, missing values depend on observed values as well as unobserved values.
4. **Blackouts**, in this certain period of time has no observations or measurements, resulting in a "gap" in the time series.

In addition to the type of missing data, we also consider the system configuration on which we conduct the experiments to assess the effectiveness and accuracy of the imputation methods under various conditions.

Methods

MICE

As a baseline model, the study uses Multivariate Imputation by Chained Equation (MICE) [5], a widely used imputation algorithm which performs multiple imputation by the application of chained equations.

GRIN

GRIN stands for Graph Recurrent Imputation Network which is a neural network model that leverages the abilities of Recurrent Neural Network, Graph Neural Network, and Graph Representation Learning [3] to impute missing values in multivariate time series dataset.

CSDI

Conditional Score based Diffusion Model (CSDI) [4] is a probabilistic imputation method which by application of conditional score-based diffusion model learns the conditional distribution in the time-series. The imputation procedure starts by filling all the missing values in the time-series by random noise. And then reverse process of conditional diffusion model de-noises the time-series which generates a plausible complete time-series. It leverages an attention mechanism which can capture the temporal and feature dependencies of the data-set.

SAITS

Self-attention based imputation for time series [2], proposed a joint-optimization approach to perform the task of missing data imputation in multivariate time series datasets. The joint-optimization approach divides the imputation process into Masked Imputation task and Observed Reconstruction Task, and the associated training loss is summation of imputation loss and reconstruction loss. To experiment this method we use PyPOTS, a toolbox designed for missing data imputation using different methods such as SAITS and BRITS.

SSSD

Diffusion-based Time Series Imputation with Structured State Space Model (SSSD) [1] is employed to model the temporal dependencies among variables, and a filtering algorithm is used to estimate unknown values by combining the observed data with predicted values from the previous time step. This method stresses specifically on different types of missing values, as mentioned in the introduction section, to extend the applicability of the method in different scenarios.

Datasets

To perform benchmark comparisons between the above-mentioned imputation methods, we run the experiments on four public real-world datasets from different domains:

1. **PhysioNet 2012 Mortality Prediction Challenge:** Contains 12000 multivariate clinical time series samples and was collected from patients in ICU(Intensive Care Unit).
2. **Beijing Multi-site Air-Quality (AQI):** Data is collected between March 1, 2013 - February 28, 2017 (48 Months) from 12 different locations in Beijing and contains 11 continuous time-series variables of hourly pollutants data.
3. **Smart grids (CER-E; Commission for Energy Regulation, 2016):** A dataset of 485 time-series observations, each comprising samples taken at 30-minute intervals, by subsetting data from smart meters used to monitor energy consumption by small and medium-sized businesses.
4. **PeMS-BAY :** This is a traffic-related dataset collected by the Performance Measurement System (PeMS) of the California Transportation Agency. The dataset consists of 6 months of data dating from January 1st 2017 to May 31th 2017, collected from 325 sensors in the Bay area.

Conclusion

In this experiment, we empirically investigate the weaknesses and strengths that affect the accuracy and efficiency of imputation algorithms. Changing the type of missing data results in significant changes in the training and testing accuracy of the algorithm. While an algorithm may produce very low mean absolute error (MAE) when applied to a dataset with a certain type of missing data, it may exhibit higher MAE when the same dataset contains a different type of missing data. Although, this study contributes to the identification of potential weaknesses and deficiencies in the algorithm and may potentially recommend modifications that can improve its accuracy, further experiments are needed to provide more comprehensive insights to recommend best algorithm being experimented and to suggest reliable methods to improve their accuracy. Our study establishes a groundwork for future research focused on creating more resilient and precise imputation algorithms for time-series data.

References

[1] Alcaraz, J. M. L., and Strodthoff, N. 2022. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*.

[2] Du, W.; Côté, D.; and Liu, Y. 2023. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications* 219:119619.

[3] Hamilton, W. L. 2020. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning* 14(3):1–159.

[4] Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34:24804–24816.

[5] Van Buuren, S., and Groothuis-Oudshoorn, K. 2011. mice: Multivariate imputation by chained equations in r. *Journal of statistical software* 45:1–67.

[6] You, J.; Ma, X.; Ding, Y.; Kochenderfer, M. J.; and Leskovec, J. 2020. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems* 33:19075–19087.