Handling missing values in data streams: An overview.

Afonso M. S. Lima[1], **Elaine P. M. de Sousa**[1]

[1]Institute of Mathematical and Computer Science (ICMC)
University of São Paulo (USP)
São Carlos – SP – Brazil

Abstract. *Missing values are a common problem in streaming scenarios, mainly due to equipment faults, network errors, and data unpredictability. This paper presents an overview of handling missing values in data streams, elucidating key concepts and summarizing recent studies that tackle this issue. It highlights limitations related to data stream requisites, concept drift exploration, and missing mechanism assumptions. Our discussion aims to indicate open issues and contribute to new research initiatives in this area.*

1. Introduction

The digital evolution has led to a significant increase in data generation, impacting all computational processes. Interconnected Internet-of-Things devices, the use of social networks and the evolution of technology in different domains generate massive amounts of streaming data, at high velocity, that have to be analyzed and processed by efficient stream mining methods [Bahri et al. 2021].

There are two main algorithmic challenges when dealing with streaming data: fast large data generation and real-time processing requirements [Bifet et al. 2023]. Any step of the knowledge discovery process (e.g., preprocessing, validation, etc.) has to consider these requisites when conceiving a new method to process data streams efficiently.

In preprocessing tasks, missing value handling is a usual necessity among stream mining problems. Failures in monitoring or data collection equipment, interruption in communication between data collectors and the central management system, failure during archiving (hardware or software), etc., are all common situations in streaming domains that motivate further studies on handling missing values for data streams. This is a complex issue, as it's necessary to consider the particularities of streaming environments, such as memory and processing limitations and data evolution (i.e., concept drift), as well as the characteristics of the missing values themselves, such as the cause of the missing value (i.e., missing mechanism) [Beyer et al. 2023].

This paper provides an overview of the current research scenario on handling missing values for data streams. We point out the key concepts in this area and compile recent papers that address this problem, identifying limitations related to data stream requirements, concept drift exploration, and missing mechanism assumption. We aim to contribute with a concise and objective discussion of the state-of-the-art solutions for missing values in streaming data, highlighting open issues and research opportunities.

The rest of this paper is organized as follows: Sections 2 and 3 outlines data streams and missing values. Section 4 summarizes related papers and discusses limitations. Finally, Section 5 presents the conclusions.

2. Data stream concepts

Data streams are unbounded sequences of multidimensional, irregular, and transient objects made available along time [Gama 2010, Bahri et al. 2021]. These characteristics imply a set of requirements that must be considered when designing and implementing methods for streaming data mining, such as one-time data read, real-time processing, limited main memory and data evolution detection [Bifet et al. 2023]. Therefore, knowledge discovery processes designed for data streams are commonly expected to: delimit the reading, preprocess data, employ fast, incremental mining methods suitable to evolving data, and validate the results [Gama 2010, Bahri et al. 2021, Bifet et al. 2023].

Figure 1 illustrates a general knowledge discovery process for data streams, considering a classification task and missing value imputation in preprocessing. First, objects from a streaming data source are delimited by, for instance, a sliding window (Win 1 in Figure 1) that retains new objects and discards old ones. Next, the objects inside the window are preprocessed to achieve processing and memory requirements and improve data quality (e.g., summarization, missing values imputation, etc). These preprocessed objects are the input of learning algorithms to generate and update the classification models. Finally, those models are continuously validated using evaluation metrics (e.g., accuracy, precision, F-measure) along with execution time and memory usage verification. These steps are continually performed as new objects are generated by the streaming data source and placed in the sliding window (Win 2 in Figure 1).

Additionally, due to the dynamic nature of data streams, any learning model should adapt to evolving data [Bahri et al. 2021, Bifet et al. 2023]. Changes in the properties of a data stream, such as distribution, recurrence, velocity, and time interval, are generally referred to as concept drift. There may also be changes in a data stream's feature space and class set. Therefore, applying concept drift detection methods to identify and track these changes and update models accordingly is essential to maintain accurate results [Hu et al. 2020, Mahdi et al. 2024].

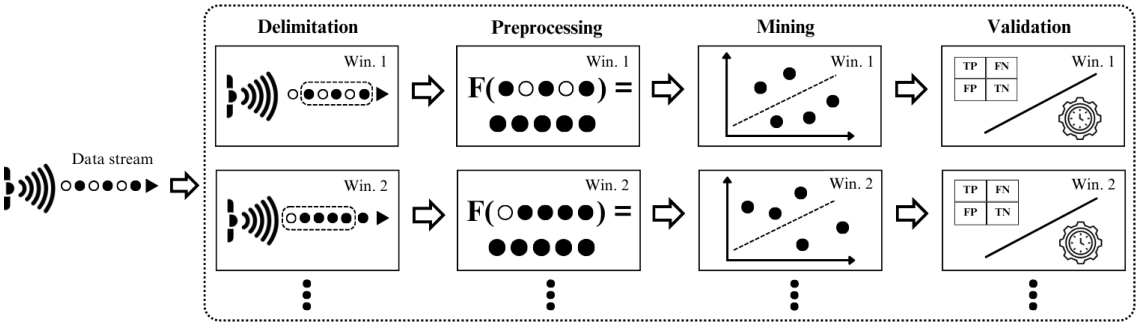

Figure 1. General steps of knowledge discovery in data streams applied to a binary classification task with missing value imputation in preprocessing: Win 1 e Win 2 (as well the following) are consecutive sliding windows; F() is an imputation method; the uncolored and the colored bullets represent objects with and without missing values, respectively

3. Missing values background

Missing values are common and unavoidable in various domains of the real world and can compromise the results of a knowledge discovery process, causing performance degra-

Table 1. Missing mechanisms summary

Missing Mechanism	Definition	Example
Missing Completely At Random (MCAR)	The missing cause is unrelated to other observed or unobserved values.	A temperature sensor failed, and the value was not generated.
Missing At Random (MAR)	The missing cause depends on other observed attributes.	Older women tend not to inform their age ("age" depends on "gender").
Missing Not at Random (MNAR)	The missing cause is related to specific information about the own missing value that is not present in the dataset.	High-income people prefer not to inform their earnings. ("income" depends on itself).

dation, data analysis problems and biased outcomes [Emmanuel et al. 2021]. Missing values may result from human mistakes when processing data, machine errors caused by equipment malfunction, respondents' refusal to answer specific questions, drop-out from studies, merging unrelated data, among other issues [Ren et al. 2023]. Thus, dealing with this problem is an essential preprocessing step for getting better results [Emmanuel et al. 2021, Ren et al. 2023].

Missing values handler methods can be divided into two principal approaches: deletion and imputation [Ren et al. 2023]. Deletion methods are simpler solutions that remove objects or attributes with one or more missing values. Imputation methods, which can usually achieve better results maintaining the original dataset size, focus on estimating and imputing values where they are missing. Applying appropriate imputation methods requires knowing, for instance, the quantity and location of missing values. More challenging is to understand the cause of the missing values, commonly known as the missing mechanism, as they can be complex and influenced by external factors. Table 1 presents the three principal missing mechanisms [Little and Rubin 2019, Emmanuel et al. 2021, Ren et al. 2023].

Most missing values methods assume the MCAR mechanism and apply simpler procedures to handle missing values (i.e., object removal), including data stream scenarios [Lin and Tsai 2020, Ren et al. 2023, Beyer et al. 2023]. However, to define that missing values are entirely unrelated to other observed attributes or specific unknown information in the dataset can compromise the quality of imputing missing values [Ren et al. 2023, Beyer et al. 2023]. Both MAR and MNAR need more robust procedures when imputing estimated values. The former requires considering the correlation between missing and not missing attributes and the latter usually requires multiple estimations to identify the external information. Although more difficult to handle, identifying all the missing mechanisms helps choose the best methods for imputing missing values for a given dataset [Ren et al. 2023, Beyer et al. 2023].

4. Missing values imputation in data streams

When applied to data streams, methods for imputing missing values must consider all the requirements and particularities of this type of data. One-time data reading, real-time processing, limited main memory and the evolving nature of data are characteristics that

Table 2. Papers approaching missing value imputation in data streams

Reference	Proposal
Fountas and Kolomvatsos (2020)	Mechanism that detects correlations between streams to input estimated values.
Sun et al. (2020)	Gamma distribution-based approach for imputing missing data in streams.
Dong et al. (2021)	Conventional methods with sliding window for imputing in streams with concept drift.
Zhang and Thorburn (2022)	Four-module system using conventional algorithms for real-time data stream imputation.
Halder et al. (2022)	Fuzzy chunk-based method for imputing in imbalanced data streams.
Liu et al. (2023)	Online algorithm using summary statistics for imputing mixed-type streaming data.
Li et al. (2023)	Message propagation network for efficient imputation at time windows in data streams.

hinder a direct application of algorithms designed for conventional databases. Since most research in data streams does not address the problem of missing values, imputing missing values in data streams remains a relatively unexplored issue.

This paper aims to make an overview type of literature review [Grant and Booth 2009], selecting research papers that approach the missing value handling problem at data stream scenarios that were published within the last four years and made available at the principal computer science digital libraries (i.e., ACM Library, IEEE Xplore and Scopus). Table 2 summarizes the accessible research.

Each related work presents a different approach to deal with imputation in data stream, mainly: conventional database imputation methods alongside a continuous processing structure [Fountas and Kolomvatsos 2020, Liu et al. 2023], summaries and statistical information of the streams [Dong et al. 2021, Zhang and Thorburn 2022], data distribution [Sun et al. 2020], fuzzy logic [Halder et al. 2022], and graph techniques [Li et al. 2023]. This scenario of few studies with diverse approaches indicates that missing value imputation in data streams is a new and currently open issue in the stream mining field that still needs further research. Furthermore, current solutions fail to fully cope with essential aspects of handling missing values in data streams, namely: data stream requisites, concept drift exploration and the missing mechanism assumption. Table 3 summarizes related work based on these aspects.

The earlier papers do not consider some data stream requirements, such as model adaptation and incremental processing of new objects, and do not explore delimitation structures (i.e., sliding windows) for these matters. Zhang and Thorburn (2022), for instance, utilize a sliding window to process the data stream, but the model is built beforehand using previously obtained data and it is never updated. Later papers meet these requirements by using both sliding windows and adaptative procedures. Still, all of them apply a batch approach to process the stream, i.e., there is

Table 3. Papers and missing values handling aspects

Reference	Data Stream Requisites	Concept Drift Exploration	Missing Mechanism
Fountas and Kolomvatsos (2020)	Not considered	Not explored	Not defined
Sun et al. (2020)	Not considered	Not explored	Not defined
Dong et al. (2021)	Considered	Explored	Not defined
Zhang and Thorburn (2022)	Not considered	Not explored	MAR
Halder et al. (2022)	Considered	Poorly explored	Not defined
Liu et al. (2023)	Considered	Not explored	MCAR, MAR
Li et al. (2023)	Considered	Not explored	Not defined

no overlapping between windows as all objects in each window shift are new ones [Halder et al. 2022, Liu et al. 2023, Li et al. 2023]. However, this approach makes it challenging to detect and adapt to certain types of concept drift, such as incremental and gradual changes.

Few papers address concept drift and conduct extensive experiments considering datasets with different types of change. Halder et al. (2022) affirm that concept drift is considered in the experiments but is limited to only one case and type of concept drift in one dataset. To truly assess a method's capacity to handle concept drift, multiple datasets that present different types of concept drift must be employed in experimental evaluation. For example, as all current methods use a batch approach, depending on the window size, the non-overlapping shift may not detect all the possible changes at the moment they occur.

Finally, few works define a missing mechanism for their datasets. Although it is possible to define it (e.g., the authors consider that there exist correlations between missing and non-missing attributes so that a MAR mechanism can be defined), this is not explicitly described in the paper [Halder et al. 2022]. Assuming missing mechanics is essential to understanding the characteristics of a dataset and to understanding a method limitation, indicating possible new development directions. Furthermore, the MNAR mechanism has not yet been explored in streaming scenarios, as it is the most challenging mechanism to handle.

5. Conclusion

We present an overview of the problem of handling missing values in data streams, compiling accessible research from the last four years. All the papers use different approaches to address this problem, showing that this issue is in its early stages and presents numerous opportunities for further research. Furthermore, we highlight the following limitations in the state-of-the-art: noncompliance with data stream requisites, naive concept drift exploration, and simple missing mechanism assumptions.

As for some potential future directions, developing a method for imputing missing values that considers online incremental processing instead of batch approaches may help address data stream challenges, such as concept drift. Extensive experimental tests on multiple types of concept drift are also necessary to assess a method's capacity to handle

such complex phenomena. Lastly, handling different missing mechanisms in data stream scenarios, including their identification and treatment, remains an open issue.

References

Bahri, M., Bifet, A., Gama, J., Gomes, H. M., and Maniu, S. (2021). Data stream analysis: Foundations, major tasks and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3):e1405.

Beyer, C., Büttner, M., and Spiliopoulou, M. (2023). Challenges for active feature acquisition and imputation on data streams. In *Proceedings of the Workshop on IAL co-located with ECML-PKDD*, volume 3470, pages 9–13, Torino, Italy. CEUR.

Bifet, A., Gavalda, R., Holmes, G., and Pfahringer, B. (2023). *Machine learning for data streams: with practical examples in MOA*. MIT press, 4th edition.

Dong, W., Gao, S., Yang, X., and Yu, H. (2021). An exploration of online missing value imputation in non-stationary data stream. *SN Computer Science*, 2:1–11.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, 8:1–37.

Fountas, P. and Kolomvatsos, K. (2020). A continuous data imputation mechanism based on streams correlation. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE.

Gama, J. (2010). *Knowledge discovery from data streams*. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Grant, M. J. and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2):91–108.

Halder, B., Ahmed, M. M., Amagasa, T., Isa, N. A. M., Faisal, R. H., and Rahman, M. M. (2022). Missing information in imbalanced data stream: fuzzy adaptive imputation approach. *Applied Intelligence*, 52(5):5561–5583.

Hu, H., Kantardzic, M., and Sethi, T. S. (2020). No free lunch theorem for concept drift detection in streaming data classification: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):e1327.

Li, X., Li, H., Lu, H., Jensen, C. S., Pandey, V., and Markl, V. (2023). Missing value imputation for multi-attribute sensor data streams via message propagation. *Proceedings of the VLDB Endowment*, 17(3):345–358.

Lin, W.-C. and Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons, Hoboken, New Jersey, USA.

Liu, W., Luo, L., and Zhou, L. (2023). Online missing value imputation for high-dimensional mixed-type data via generalized factor models. *Computational Statistics & Data Analysis*, 187:107822.

Mahdi, O. A., Ali, N., Pardede, E., Alazab, A., Al-Quraishi, T., and Das, B. (2024). Roadmap of concept drift adaptation in data stream mining, years later. *IEEE Access*, 12.

Ren, L., Wang, T., Seklouli, A. S., Zhang, H., and Bouras, A. (2023). A review on missing values for main challenges and methods. *Information Systems*, page 102268.

Sun, Z., Zeng, G., and Ding, C. (2020). Imputation for missing items in a stream data based on gamma distribution. In *International Conference on Smart Computing and Communication*, pages 236–247. Springer.

Zhang, Y. and Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, 128:63–72.