

Challenges for Active Feature Acquisition and Imputation on Data Streams

Christian Beyer^{1,*}, Maik Büttner¹ and Myra Spiliopoulou¹

¹Otto-von-Guericke-University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany

Abstract

Two popular methods for dealing with missing feature values are active feature acquisition as well as imputation. Both methods often require an understanding of a feature's relationship to the target variable as well as to all the other features. Developing such an understanding is time-consuming and challenging in a static setting, but becomes much more complicated in a data stream scenario. Additional challenges are concept drift, feature drift, incorporating feature costs, dealing with complex types of missingness, and the need for imputation models that can be updated efficiently. In this work, we will discuss these challenges as well as challenges that appear downstream when devising stream-applicable solutions. The goal is to provide a current overview and inspire discussion as well as further research in this field.

Keywords

active learning, data streams, imputation, active feature acquisition

1. Introduction

Many machine learning models can only be trained and create predictions if all the features of an instance are available. This imposes a need for methods to replace missing feature values during training and testing. Imputation is the most popular approach to deal with missing values, where the missing feature values are estimated using heuristics and models, that either rely on a feature's distribution or its relationship to other features [1]. Another approach to deal with missing values is Active Feature Acquisition (AFA), where the real feature values can be purchased from a costly oracle under budget constraints [2]. An oracle could be a costly subject matter expert that has to be inquired or a lab test that has to be done. Both approaches have different advantages and disadvantages, see Table 1. This makes them applicable in different scenarios and often it would make sense to use a mix of both methods. For example, if an instance has a feature missing, that is strongly correlated with another feature that is available, then using imputation might be a good strategy. On the other hand, if the missing feature cannot be predicted well by available features, it might require a costly lab test or the opinion of a subject matter expert to determine the real feature value. This hypothetical scenario shows, that knowledge about the features and their relationship to one another and their relationship to the target variable is needed. The inter-feature relationships are needed in order to build proper models for imputation as well as to guide AFA methods toward purchases of features


IAL@ECML-PKDD'23: 7th Intl. Worksh. & Tutorial on Interactive Adaptive Learning, Sep. 22nd, 2023, Torino, Italy

✉ christian.beyer@ovgu.de (C. Beyer); maik.buettner@ovgu.de (M. Büttner); myra@ovgu.de (M. Spiliopoulou)

🆔 0000-0001-8604-9523 (C. Beyer); 0000-0002-1828-5759 (M. Spiliopoulou)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

that cannot be inferred from available ones. The relationship to the target variable is needed to know if a certain feature is even relevant to the task at hand, which is often called the *merit* of a feature. If it is not relevant then we can safely remove it, as the feature value should not influence the final decision. In case it is relevant, then we should consider purchasing it if it cannot be imputed with high confidence. In contrast to a static setting, in a data stream all these relationships might be subject to change and imputation models that map the inter-feature relationship need to be updated often and in a timely manner. These and subsequent challenges will now be discussed in more detail.

Approach	Advantages	Disadvantages
Imputation	<ul style="list-style-type: none"> - fast compared to AFA - no cost - can usually be applied to the whole data set 	<ul style="list-style-type: none"> - biased estimates - can be wrong - requires representative data - imputation method has to match data properties - Most methods only designed for MCAR
Active Feature Acquisition	<ul style="list-style-type: none"> - real values - no need for representative data 	<ul style="list-style-type: none"> - costly - can be slow - can usually only be applied to a fraction of the data set

Table 1
Comparison of imputation and AFA in a static setting

2. The Challenge of Dealing with Different Types of Missingness

Almost all stream-related publications deal with one type of missingness which is missing completely at random (MCAR) [1]. This makes other types of missingness an under-researched area. MCAR means that the missingness of a certain feature is independent of factors within the data set and cannot be explained by outside factors as well. Though easy to model, it is also the least likely case to be encountered, as in most cases the missingness either depends on the variable itself or on variables inside or outside the data set. An example of the former would be, if older people were less likely to state their age, and an example of the latter would be if people of a certain gender would tend to skip certain questions. The challenges are to detect which type of missingness we are confronted with and develop stream-applicable methods that can handle missingness apart from MCAR. In [3] the authors considered a special scenario where all the features of an instance are missing and are purchased iteratively and while their approach is designed for streams, it seems very inadequate to deal with drift among the features. Another interesting solution is presented in [4], where the authors propose a deep ladder imputation network that can handle any kind of missing data and also deal with high degrees of missingness. Unfortunately, it is again ill-suited for streaming data containing drift. Real-time induction and updating of the model are additional open challenges.

3. The Challenge of Dealing with Drift

In an incomplete stream two types of drift may occur: feature drift and concept drift. Feature drift occurs if the distribution of a feature changes or if the relationship of the feature to the target variable changes [5]. Feature drift can occur abruptly, gradually, or shifting and necessitates an update of the imputation models as well as the *merit* estimates used by AFA. In order to address potential feature drift, we first have to detect it. This constitutes another challenge in itself as drift detection algorithms are specialized in detecting certain types of drift and there is no free lunch [6]. Once feature drift has been detected it is important to forget outdated information and to update the imputation models and *merit* estimates. There are also simple solutions like windowing techniques [7] but windows of static length have the disadvantages that we might miss moments of feature drift and apply outdated models for a while or that the imputation and *merit* estimates are subpar because the available training data is artificially restricted by the window length.

A temporal change in the distribution of the target variable or a change in the relationship of the target variable to other features is called concept drift [6, 8]. Concept drift does not affect imputation directly, as the target variable is usually not used, but a change in the target variables distribution could exacerbate the problem of biases in the imputed values. For example, if we consider an imputation method that always replaces a missing feature with the feature mean and a highly skewed data set, where the majority class often has values around the feature mean associated with it, then we will produce only a few prediction errors. If concept drift happens so that the minority class is more prominent around that feature's mean then we might introduce a lot of errors, because the predictions will now favor the minority class. The problems for AFA are more obvious as a feature's *merit* is supposed to inform us how valuable it is for solving the task at hand, which often means how well it helps in separating the classes [9]. If classes now start to overlap or change abruptly, then *merit* estimates will need to be updated and changed accordingly. This again necessitates first of all that we recognize when drift happens.

Deep-Learning-based approaches are becoming increasingly popular as they achieve high performances [3, 4, 10] but are highly susceptible to the issue of drift. Their need for a lot of representative data and computational time to adapt to new concepts makes them ill-suited for such scenarios.

4. The Challenge to Induce Imputation Models in Near Real Time

If we consider the detection of feature and concept drift solved then we are still left with the need for imputation models that can be updated in almost real-time. This excludes or makes the application of several prominent imputation methods like MICE [11] and methods based on deep neural networks [4] much harder because they require multiple runs over the same training data which can be very time-consuming. The structure of inter-feature relationships could be modeled with a Bayesian network [12] and used for imputation but online versions have shown to be subpar to static versions [13] and it is also challenging to adapt Bayesian Networks to different types of drift [14], especially shifting drift. One proposed solution to

make algorithms designed for static data viable on data streams, is the usage of windowing techniques [7] to restrict training data. However, these cannot always be applied especially when we want to employ deep learning-based methods which promise high imputation quality. Knowledge about the inter-feature relationships would also be of high value in an AFA setting. One drawback of stream applicable *merit* estimates [9] is, that these estimates are independent for each feature and therefore ignore all feature-to-feature relationships which could potentially be exploited. It could therefore happen that we inquire a costly oracle to provide a feature value that could have been predicted very well by other features that were available.

5. The Challenge of Dealing with Feature Costs

Features can have varying costs, for example, measuring a patient’s temperature requires less costly materials and less skill than running an MRI. Costs do not have to be monetary. They can also describe the time or expertise required to acquire a feature. These varying costs might introduce an additional bias towards cheaper features in the selection process of feature acquisitions when the budget is a further constraint to consider. Such biases worsen the problem of the trade-off between exploration and exploitation whereby neglecting to purchase specific features due to their inhibiting cost might delay the detection of new feature concepts and inter-feature relationships. However, incorporating feature costs has been shown to improve predictive performance in static settings under budget constraints [15]. Our early work on data streams supports this notion [16]. We pointed out that it requires more complex *merit* functions. These should take the inter-feature relationships into account. Feature costs are also a motivation to tackle the challenge of combining AFA and imputation in an intelligent manner so that the budget is only spent to purchase feature values that cannot be imputed well. In the case of missing labels, queries for both labels and features may be combined to allow learning agents themselves to decide on the trade-off of prioritizing training imputation models, training prediction models, adapting models to drifts, and saving budget.

6. Conclusion

In this short work, we motivate challenges that impede the application of imputation and AFA methods on data streams, especially when we want to apply them in a joint framework. The main challenges are:

- Need for imputation models that can handle any type of missingness
- Need for fast, high-performance imputation models that can be updated incrementally
- Need for a general feature and concept drift detection
- Modelling the inter-feature relationship in the face of different kinds of drift
- Need for AFA methods that take feature costs into account

We hope this work will encourage discussion, as well as future research that addresses these challenges.

References

- [1] W.-C. Lin, C.-F. Tsai, Missing value imputation: a review and analysis of the literature (2006–2017), *Artificial Intelligence Review* 53 (2020) 1487–1509.
- [2] B. Settles, Active learning literature survey, Technical Report 1648, University of Wisconsin-Madison Department of Computer Sciences, 2009.
- [3] M. Kachuee, O. Goldstein, K. Kärkkäinen, S. Darabi, M. Sarrafzadeh, Opportunistic learning: Budgeted cost-sensitive learning from data streams, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019.
- [4] E. Hallaji, R. Razavi-Far, M. Saif, Dlin: Deep ladder imputation network, *IEEE Transactions on Cybernetics* 52 (2021) 8629–8641.
- [5] J. P. Barddal, H. M. Gomes, F. Enembreck, B. Pfahringer, A survey on feature drift adaptation: Definition, benchmark, challenges and future directions, *Journal of Systems and Software* 127 (2017) 278–294.
- [6] H. Hu, M. Kantardzic, T. S. Sethi, No free lunch theorem for concept drift detection in streaming data classification: A review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (2020) e1327.
- [7] W. Dong, S. Gao, X. Yang, H. Yu, An exploration of online missing value imputation in non-stationary data stream, *SN Computer Science* 2 (2021). doi:10.1007/s42979-021-00459-1.
- [8] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under concept drift: A review, *IEEE Transactions on Knowledge and Data Engineering* (2018) 1–1.
- [9] L. Yuan, B. Pfahringer, J. P. Barddal, Addressing feature drift in data streams using iterative subset selection, *ACM SIGAPP Applied Computing Review* 19 (2019) 20–33.
- [10] J. Kossen, C. Cangea, E. Vértés, A. Jaegle, V. Patraucean, I. Ktena, N. Tomasev, D. Belgrave, Active acquisition for multimodal temporal data: A challenging decision-making task, *Transactions on Machine Learning Research* (2023).
- [11] S. Van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in r, *Journal of statistical software* 45 (2011) 1–67.
- [12] R. Howey, A. D. Clark, N. Naamane, L. N. Reynard, A. G. Pratt, H. J. Cordell, A bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships, *PLoS Genetics* 17 (2021) e1009811.
- [13] P. Ratnapinda, M. J. Druzdzal, Learning discrete bayesian network parameters from continuous data streams: What is the best strategy?, *Journal of Applied Logic* 13 (2015) 628–642.
- [14] Q. Meng, Y. Wang, J. An, Z. Wang, B. Zhang, L. Liu, Learning non-stationary dynamic bayesian network structure from data stream, in: 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), 2019, pp. 128–134. doi:10.1109/DSC.2019.00027.
- [15] O. Kaminska, T. Klonecki, K. Kaczmarek-Majer, Feature selection in bipolar disorder episode classification using cost-constrained methods, in: *Artificial Intelligence in Medicine*, Springer International Publishing, 2023. To be published.
- [16] M. Büttner, C. Beyer, M. Spiliopoulou, Reducing missingness in a stream through cost-aware active feature acquisition, in: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2022, pp. 1–10.