Imputation of missing data in a domain adaptation context

Laetitia Chapel, Institut Agro - IRISA
Romain Tavenard, Université Rennes 2 - LETG

Internship
Expected starting date: March/April 2024
Possibly followed by a PhD

Key-words Machine learning, unsupervised and supervised domain adaptation, missing data imputation, time series.

Context. AI methodologies typically depend on extensive datasets that may be tainted by noise, missing values, or can be collected in heterogeneous yet related environments. Data with missing values are ubiquitous in many applications; they can be due to equipment failure, incomplete information collection (e.g. clouds in the remote sensing case) or inadequate data entry for instance. Nevertheless, conventional learning algorithms often assume that the data are *complete* and *independent and identically distributed*, that is to say they have been drawn randomly from a single distribution.

Data imputation aim at substituting missing data by *plausible* values [1], e.g. by filling them by the value of the nearest sample or by imputing with some relevant statistics. The imputation can have a high impact on performances of the learning task at hand, leading to biased results or degraded performances [2, 3]. Most of the imputation methods rely on some *(completely) missing at random* assumption [4] and with no pattern between the missingness of the data and any values. More challenging scenario deal with random block missing or blackout missing [5], in which blocks of information are missing and where the structure of block-wise missing data should be further taken into consideration.

On the other hand, the outcomes generated by AI play a crucial role in monitoring and comprehending environmental phenomena through the resolution of various tasks, including but not limited to:

- land cover and land use mapping, that can then further being used for urban planning, agriculture management, or identifying illegal land use activities for instance;
- crop yield prediction, in order to ensure food security, economic stability, and sustainable agricultural practices;
- wildlife conservation, in which wildlife habitats, migration patterns and population changes can be evaluated at a large scale;
- fisheries control, to identify, measure and check the aquatic resources that are harvested, aiming to protect over-exploited aquatic species.

In practice, the data are often collected on different yet related domains, offering the potential to enhance the generalization capability of the learning algorithm. For instance, in Earth observation, and especially for land cover mapping applications, the differences in weather, soil conditions or farmer practices between study sites are known to induce temporal shifts that can be corrected to enhance task performance. For predicting crop yield, the variability under changing climates and severe weather events have to be taken into account when considering data from the past to predict the evolution of the yield. *Domain adaptation* [6, 7] aims to transfer knowledge from one domain to another and has demonstrated significant enhancements in classification or clustering tasks when domain shifts are carefully managed.

Scientific objectives and expected achievements. The aim of the internship is to study the potential of data imputation method within the context of domain adaptation. Existing approaches mostly tackle missing values within an inferential framework, wherein they are replaced with values derived from dataset statistics, relying on robust parametric assumptions. However, when a shift exists between the datasets, this strategy becomes inadequate. Instead, we propose to address imputation and learning tasks concurrently, introducing the additional complexity that the data may originate from different domains. The primary objectives of the internship are as follows: i) make an extension literature review on data imputation, with a focus on random missing block scenario ii) theoretically analyze the impact of domain shifts on the learning task, akin to the framework established for domain adaptation in a classification context (cite [8]); iii) introduce novel imputation schemes in heterogeneous environments by aligning distributions in a preliminary step and subsequently applying learning tasks (e.g., supervised learning). This last step will be a step behind the definition of an integrated framework for imputation and learning in heterogeneous environments, that would be the topic of a PhD. Special attention will be given to handling time series data, where the challenge is heightened due to the need to account for temporal correlations between observations. This will allow considering the *blackout missing* case as a special case of time series prediction.

The research directions will explore optimal transport-based solutions, known for their success in imputing missing values [9] and aligning distributions in a domain adaptation context [10], especially when dealing with temporal data [11].

From an applicative view point, remote sensing data, that are known to provide data with (spatial or temporal) shifts, will serve as a playground to evaluate the literature and validate the proposed methods.

Research environnement/Location/Supervision. The research will take place within the MALT research group from IRISA in Rennes. MALT is a newly created team whose focus is on machine learning in structured environments. The internship will be directed by Romain Tavenard and Laetitia Chapel.

Candidate profile Applicants are expected to be graduated in mathematics/statistics and in computer science and/or machine learning and/or signal & image processing, and show an excellent academic profile. Beyond, good programming skills are mandatory.

Application procedure Send a resume to Laetitia Chapel (laetitia.chapel@irisa.fr) and Romain Tavenard (romain.tavenard@univ-rennes2.fr). Potential candidates will be contacted for interview. Feel free to contact us for any question.

References

[1] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, “A gentle introduction to imputation of missing values,” *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[2] T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. V. Torne, E. Sala, P. Lio *et al.*, “Classification of datasets with imputed missing values: Does imputation quality matter?” *arXiv preprint arXiv:2206.08478*, 2022.

[3] Z. Zhang, X. Xiao, W. Zhou, D. Zhu, and C. I. Amos, “False positive findings during genome-wide association studies with imputation: influence of allele frequency and imputation accuracy,” *Human Molecular Genetics*, vol. 31, no. 1, pp. 146–155, 2022.

[4] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” *Journal of Big Data*, vol. 8, no. 1, pp. 1–37, 2021.

[5] F. Xue and A. Qu, “Integrating multisource block-wise missing data in model selection,” *Journal of the American Statistical Association*, vol. 116, no. 536, pp. 1914–1927, 2021.

[6] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, “A brief review of domain adaptation,” *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[7] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, *Advances in domain adaptation theory.* Elsevier, 2019.

[8] ——, “A survey on domain adaptation theory: learning bounds and theoretical guarantees,” *arXiv preprint arXiv:2004.11829*, 2020.

[9] B. Muzellec, J. Josse, C. Boyer, and M. Cuturi, “Missing data imputation using optimal transport,” in *International Conference on Machine Learning.* PMLR, 2020, pp. 7130–7140.

[10] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, “Joint distribution optimal transportation for domain adaptation,” *Advances in neural information processing systems*, vol. 30, 2017.

[11] F. Painblanc, L. Chapel, N. Courty, C. Friguet, C. Pelletier, and R. Tavenard, “Match-and-deform: Time series domain adaptation through optimal transport and temporal alignment,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2023, pp. 341–356.