Low rank tensor approximation techniques for up- and downdating of online time series clustering

A Data Management Plan created using DMPonline.be

Creators: Wannes Meert, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Wannes Meert, n.n. n.n.

Data Manager: Wannes Meert, n.n. n.n.

Project Administrator: Wannes Meert, n.n. n.n.

Grant number / URL: G0A9923N

ID: 200724

Start date: 01-01-2023

End date: 12-12-2027

Project abstract:

Time series clustering is used to analyse datasets generated by sensors, IoT devices, digital networks, etc. It allows to discover structure and patterns by grouping similar behavior for tasks as explanations, forecasting, and anomaly detection. Current methods struggle to deal with present datasets and needs by users. First, datasets have grown to such scale that clustering requires millions to trillions of comparisons, which one is unable to perform with current methods. Second, users increasingly expect to include human supervision to steer the clustering: semi-supervisedinstead of unsupervised clustering. Third, it is not always possible to store or process all historical data to restart clustering. When new data arrives an online method is prefered that builds upon the earlier acquired results.

In this project, we develop theory, algorithms, and implementations for clustering online time series with expert info by relying on tensors, while reducing the computational and storage needs. To deal with groups of time series we switch to similarity tensors built via theoretically supported techniques such as adaptive cross approximation (ACA). Tensor fusion allows us to add expert knowledge. Cluster information will be revealed through the use of low-rank factorizations, which we can up- and downdate efficiently to get online. Moreover methods such as ACA offer error bounds on the approximation, which is important to guarantee a consistent clustering, even on large datasets.

Last modified: 27-06-2023

Low rank tensor approximation techniques for up- and downdating of online time series clustering
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

The research will generate manuscripts (papers in preparation) and corresponding software. The software will be custom-made software written in, e.g., Matlab, Python, Fortran, or C implementing the methods described in the project proposal. The numerical experiments performed will be described in detail within the software repository. The corresponding input data will also be made available in data files or in the form of scripts that generate this input data. This means that the code as well as the corresponding explanation will be made available as text files.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

a. The responsible person will be Raf Vandebril, Department of Computer Science, KU Leuven.

b. Internal reports as well as papers will be made available using the Lirias system (KU Leuven).

For the short as well as for the long-term, the other relevant research data, such as software and the corresponding user guides and input data, will be stored on a secure NetApp-based storage solution at the Dept. of Computer Science (KU Leuven). Secure backups are automatically stored at a second location at KU Leuven, so loss of data is minimized.

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

There is no reason to deviate from the minimum preservation of 5 years. Following KU Leuven policy, storage is foreseen for at least 10 years.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

There are no such issues. No data will be collected in this project.

The challenges in this project are motivated by real-world, industrial cases in other projects. Although we do not foresee this, if a real-world use case would be deemed valuable to validate this project and a need for data sharing arises, we will follow the required procedures for data sharing. For personal data, an approval request will be submitted to the ethics committe at KU Leuven (PRET/SMEC) and the approval will be communicated with FWO. For proprietary data, a proper license agreement will be set up with the assistance of Leuven Research & Development (LRD, KU Leuven).

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

There are no other relevant issues.

Low rank tensor approximation techniques for up- and downdating of online time series clustering
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
UCR Time Series Classification Archive	Collection of time series benchmarks freely usable for research (and most also for commercial usage).	Reuse existing data	Digital	• Observations • Experimental • Compiled/aggregated data	• *.csv* • *.ts* • *.arff*	<100GB (individual datasets from this benchmark suite are <1GB)	NA
Time Series Classification Benchmarks	Collection of time series benchmarks freely usable for research (and most also for commercial usage).	Reuse existing data	Digital	• Observations • Experimental • Compiled/aggregated data	• *.csv* • *.ts* • *.arff*	<100GB (individual datasets from this benchmark suite are <1GB)	NA
Source code	Code that implements the algorithms proposed during this project.	New	Digital	• Source code	• Python • Fortran • C • Matlab	<100MB	NA

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

UCR Time Series Classification Archive
@misc{UCRArchive2018,
title = {The UCR Time Series Classification Archive},
author = {Dau, Hoang Anh and Keogh, Eamonn and Kamgar, Kaveh and Yeh, Chin-Chia Michael and Zhu, Yan and Gharghabi, Shaghayegh and Ratanamahatana, Chotirat Ann and Yanping and Hu, Bing and Begum, Nurjahan and Bagnall, Anthony and Mueen, Abdullah and Batista, Gustavo, and Hexagon-ML},
year = {2018},
month = {October},
note = {\url{https://www.cs.ucr.edu/~eamonn/time_series_data_2018/}}
}

Time Series Classification Benchmarks
https://www.timeseriesclassification.com

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

The datasets included in these benchmarks are not related to humans or are fully anonymized (and not sensitive).

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

In this project we will not collect data.
The challenges in this project are motivated by real-world, industrial cases in other projects. Although we do not foresee this, if a real-world use case would be deemed valuable to validate this project and a need for data sharing or collection arises, we will follow the required procedures. For personal data, an approval request will be submitted to the ethics committe at KU Leuven (PRET/SMEC) and the approval will be communicated with FWO.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

The computer code generated (i.e., the algorithms) have potential for commercial valorization. This is independent of the datasets used to validate the methods.
The code will be released as open-source where valuable for the research objectives. When potential for valorization is identified the benefits of a proprietary license will be analyzed first together with KU Leuven Research & Development.
When releasing the code as open-source, this will be as part of the DTAIDistance open-source package that uses the Apache Public License v2 (copyright fully at KU Leuven).

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

The benchmark datasets are owned by third-parties but are only used for validating the algorithms produced in this project. These datasets have the appropriate license to allow for such research use. We will not redistribute the benchmark datasets or any derived results.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

The code will be packaged as a Python package with a full documentation about how to install the package and each public-facing method documented (i.e, the API). Standard Python documentation practices will be used: a README file for intro and licensing info; docstrings per method aggregated using Sphinx; example Jupyter notebooks; PyTest Unit tests.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

The source code will be following Python distribution (PyPI) and Github/Zenodo principles: keywords, urls, etc. are included in the setup.cfg file and discoverable through PyPI, Github and Zenodo entries.

3. Data storage & back-up during the research project

Where will the data be stored?

Code is stored and versioned using the KU Leuven Gitlab repository server.
The datasets used for experiments will be stored on the Dept CS in-house storage solution (NetApp). Additionally, datasets are available online in the original repositories.
A snapshot of the code, experiments and data will be created for every publication.

How will the data be backed up?

The Gitlab repository server and the NetApp data storage at CS are both backed up using the KU Leuven backup services.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The Gitlab repository server and the NetApp data storage are only accessible (via authentication) to members of the project.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The Gitlab repository server is offered by KU Leuven.
The Dept CS NetApp datastorage is covered by the budget of this project. After this project ends this cost is taken over by the research group(s)' reserves.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All code generated will be stored in the KU Leuven Gitlab repository server and on the Dept CS NetApp storage (the latter guaranteed for 10 years).
Additionally, the open-source code will also be available in public repositories such as Github and Zenodo.

Where will these data be archived (stored and curated for the long-term)?

The Dept CS NetApp storage solution (>10TB) will retain snapshots for every publication (source code, experiments, manuscript) for at least 10 years.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The Dept CS NetApp longterm storage is covered by the research group's reserves. For this project, this is expected to be <100 euros / year (this does not include costs for computation which is only relevant during the project).

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, ...)
- No (closed access)

All publications will be made available as Open Access in Lirias.
All open-source code will be made available via Gitlab (and additionally Github/Zenodo public repositories).
All proprietary code (if deemed necessary for a valorisation track) will not be made available outside of the research group.
The datasets are third party and already available under an open-source license (or similar for the purpose of the research).

If access is restricted, please specify who will be able to access the data and under what conditions.

Proprietary code will only be available to team members of this project. If a followup valorisation trajectory is defined, members to that track will also be given access. This is to be defined later.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights
- No

All results of this research can be shared without any restrictions, with the exception of code that will be used in a valorisation track.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Source code will be shared using public and popular repository servers such as Github. Additionally, the code is also shared using the EU's Zenodo service (using the Github integration).

When will the data be made available?

Open-source code will be shared after publication of the research (possibly earlier if allowed by the publication venue).

Which data usage licenses are you going to provide? If none, please explain why.

Code that will be integrated with the existing DTAIDistance toolbox will follow the Apache Public License v2.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

We will not distribute any datasets as this project does not generate data.
Source code will receive a DOI using the Zenodo service.

What are the expected costs for data sharing? How will these costs be covered?

There is no cost, we will make use of publicly offered services.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Prof. Raf Vandebril

Who will manage data storage and backup during the research project?

Dr. Wannes Meert

Who will manage data preservation and sharing?

Dr. Wannes Meert

Who will update and implement this DMP?

Dr. Wannes Meert

Low rank tensor approximation techniques for up- and downdating of online time series clustering
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Low rank tensor approximation techniques for up- and downdating of online time series clustering
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable