Communication Protocol

Ismail Wahba

July 2020

Abstract

This document is a communication protocol for the master thesis. It includes a dated log for all the communication events that have happened through the course of the master thesis. This includes meeting minutes, email discussions, questions that come up in between periodic appointments and also a documentation of their statuses and their answers.

5th October 2020, Reply to Follow up Email and questions about libraries

Dear Ismail Wahba.
Great overview, very promising. I had hoped that WEASEL, BOSS and TEASER were also in pyts or sktime, but 16 classifiers in sktime sounds very promising. From what I read, sktime has the most extensive functionalities.
To place information in context:

- which of these repos contain algorithms by Bagnall and team?
- ... by Keogh and peers?
- ... eTSCA (by any author)?

All todos are as we discussed. There is one more todo:

- Specify what is meant by "incremental learning"

After all discussions we had, this specification is a must, so that the objectives of the analysis are clear.
All the best, Myra Spiliopoulou

5th October 2020, Follow up Email

Dear All,
Please find below the updates for the period between 25 Sept - 4 Oct:

- Done
  - Read the papers of TEASER and InceptionTime (thank you Noor for sharing it with me)
  - Investigated multiple github repositories with different implementations of TSCA algorithms:
    - sktime
      - The main repository, it is an extension of sklearn for TS data
      - Contains 16 different classifiers and data importing capabilities
      - Allows for the creation of pipelines and ensembles
    - pyts
      - Connected the UCR and UAE data archives (the biggest 2 archives)
      - Repository with good utils for transformation and approximation
      - Implementation of 6 different classifiers
    - sktime-dl
      - Repository for deep learning algorithms adjusted for TSC
      - Contains 10 different deep learning classifiers including (InceptionTime, CNN and ResNet)
    - SFA_Python
      - Repository for eTSCA TEASER, univariate (WEASEL, BOSS and BOSSVS) and multivariate (WEASEL+MUSE)

  There is no repository offering everything, but sktime and pyts seem to be well integrated, I had a quick play with both. I also tried downloading a dataset using pyts and then importing it in sktime and succeeded.
  I am also hoping that sktime-dl will be easy to use with them as it was developed by the same team of sktime.

- To-do
  - Write a proposal based on our last call all together
    - 3 categories (conventional TSCA, eTSCA with full TS learning, eTSCA with incremental learning)
    - Given a dataset, the framework will do the analysis and recommend the best algorithm(s) to use
    - The framework should provide insights on the earliness and performance for the algorithms on the dataset
  - Write a time-plan for the coming milestones
  - Short-list the algorithms that will be used

Best Regards, Ismail Wahba

24^{th} September 2020, on the subtopic "incrementality" - 5 The Call

- Incrementality does not mean that the number of observations for an instances increases by time, but it is related to how we want to show the data to the classifier
- We will not an early classification in the strict sense of "incremental learning", but we will create a testbed where we simulate an earliness context
- We want to consider 2 criteria in our experiment : earliness and quality
- There are 2 types of early classification algorithms: ones that see the whole series and others that see only parts from the beginning
- We are not doing event classification
- Check TEASER algorithm, as it doesn't need to see the whole time series in learning
- Our goal is nor creating incremental learners, we are evaluating existing learners on how early they can classify. By learning "offline" on the data while revealing more and more data
- Desirable goal is to try to optimize the process, since going offline, learning and then testing will be very costly. Not by optimizing every algorithm, but by optimizing the framework
- We create a wrapper around a group of algorithms, this wrapper will run the algorithms on the dataset and recommend the best one based on how every algorithm perform on the dataset
- The dataset is not a parameter of the recommendation, for example we will not recommend algorithms based on similarity between datasets
- The parameters used for the recommendation: earliness, quality (accuracy) and a metric of overall performance (Pareto optimization)
- Use 2 families: Conventional and early time series classification algorithms
- For conventional classifiers we will chop the data
- For early classifiers that see the whole time series we chop the data
- For early classifiers that see part of the data we will not do chopping, because it doesn't see the whole data anyway (TEASER family)
- To have comparable results between TEASER and other algorithms that we will do the chopping for, we need to have comparable measure. Because TEASER calculates earliness for each time series seperately, but other algorithms earliness is based on the chopping

- Differentiate between dependent instances (instances belong to same object) and independent instances (panel data)

23rd September 2020, on the subtopic "incrementality" - 4

Dear Ismail Wahba.
Today I spoke with Noor: indeed, a live discussion clarifies a lot. We will clarify all.
Till tomorrow,
Myra Spiliopoulou

23rd September 2020, on the subtopic "incrementality" - 3

Dear Prof Spiliopoulou,
Thank you for the reply.
Yes it helps regarding the first point of connecting/transforming TSCA to E-TSCA.
As for the incremental property of algorithms, I believe it will become clearer during our discussion on Friday.
Best Regards,
Ismail Wahba

23rd September 2020, on the subtopic "incrementality" - 2

Dear Ismail Wahba.
This mail goes to you and also to Vishnu and Noor, because Vishnu told me that you exchange with them and have difficulties with item b) below.
We have a meeting on Friday, where we can discuss this. I think that the subject is much simpler than it sounds:
There are algorithms for time series classification (TSCA) and methods for early time series classification (E-TSCA).
What must be done to transform an arbirary TSCA into an E-TSCA?
One way to go is that you pick two state of the art TSCAs and propose an E-extension for each one.
Does this help?
Best regards,
Myra

23rd September 2020, on the subtopic "incrementality" - 1

Dear Prof Spiliopoulou,
Since our last call I have been investigating the two topics we discussed

- Early Timeseries Classification as one of the motivations for our research
- How can we transform a normal TSCA to an incremental one

As for the first point, I have made myself more familiar with the idea of early classification, although while reading I came to realize more and more that the technique itself is different than TSCA.
My problem was with the second point. Whenever I found something with incremental learning, it is discussing online learning (data streams), which I believe is not related to our topic.
I couldn't find anything discussing incremental learning in an offline context except one paper. (Losing, Viktor, Barbara Hammer, and Heiko Wersing. "Incremental on-line learning: A review and comparison of state of the art algorithms." Neurocomputing 275 (2018): 1261-1274.)
In the context of the paper, I understood that incremental learning is more like mini batches learning. We divide the dataset into N batches. Starting with the first batch (subset of instances) we learn a model, then for each new batch we learn a new model which is based on the previous model and the batch at hand (subset of instances). Then the classification is done using only the last model.
But I think this is not the same setting for our research, as we want to learn a model on few time points (features) for all instances and then make the model learn more by incrementing more time points (features) to it.
I feel like I am stuck now and I am seeking guidance.
I am not sure if that really is the incremental learning technique that we should apply ?
and how to connect the incremental learning topic to our research ?
I appreciate your help
Best Regards,
Ismail Wahbha

26th August 2020, Reply to mail of Aug 23

Dear Ismail Wahba.
I hope I did not miss much from the communication. It seems you have some difficulties with WEASEL, you need time for it.
Concerning ANOVA, I suggest that you find a tutorial with code and some hands-on exercise, nothing very complicated, so that you can walk through it and understand how it works.

Meeting: Sept 1, 16:30 via skype?
Best regards, Myra Spiliopoulou

26th August 2020, Reply to mail of July 27

Dear Ismail Wahba. I have traced two documents in overleaf, the communication protocol and the literature overview. if there are more masterfiles, I did not findi them...
The literature work is progressing well. You do not need to go into deep details of the papers you read, but you must understand

- how they work and
- how you should apply them when you start "shortening" the time series.

In that sense, below are a few remarks on the papers you ask questions about:

- ☑ Paper by Bagattini et al: I didn't understand the part of handling missing values from the timeseries
  - ☑ They use a symbolic representation, where the timestamps are mapped to an ordering scheme (as in SAX). Then, they build the s-shapelets. For your thesis, please consider whether the construction of this symbolic representation and the generation of the shapelets demands that the algorithm has seen the whole timeseries. If yes, then each time you shorten the timeseries you need to derive the symbolic representation anew and then generate the shapelets. An implication is that the size of the alphabet (for the symbolic representation) may shrunk substantially, unless all symbols are equiprobable anywhere in the timeseries.
- ☐ Paper by Baydogan et al: I did't understand how did they discretize the CPE and created the histograms
  - Unfortunately, I cannot get this paper from outside the faculty. But my suggestions for the previous paper may also apply here: do they build the histogramms over the whole time series? If yes, then when the time series is shortened, the histogramms change.
- ☑ For now, do I have to read the details about complexity and understand the details of the algorithm ? or is it enough to understand how it works ?
  - Issues (a) and (b), as said in the beginning.

26th August 2020, Follow up emails after vacation and Parental Leave

Dear Ismail Wahba. Sorry for being so slow, I am now going to catch up. I marked the overleaf project. It is fine.

- ☑ I recommend that you continue filling it, but most recent communication first.

all the best, Myra Spiliopoulou

27th July 2020, Communication Protocol and Literature Overview Document

Dear Prof Spiliopoulou, Here is my update for the week 21 jul - 27 jul: The link for the overleaf repository: https://www.overleaf.com/read/frpfdfwhgkgc Actions from last week:

- ☑ Finish reading the papers of Ulf Leser and Bagattini et al
  - ☑ Finished reading the papers by 1)Ulf Leser 2)Bagattini et al. and 3)Baydogan et al. I also added them to the literature overview document
- ☐ Check literature for correlated time series it might have some points useful to our problem
- ☑ Collect broader literature regarding Multivariate Time Series Classification
  - ☑ started reading the WEASEL paper of Ulf leser 2017
- ☑ Collect more Review Papers
  - ☑ Included the literature review of the paper by Ulf leser into the review folder

Actions till next week:

- Check literature for correlated time series it might have some points useful to our problem
- Collect broader literature regarding Multivariate Time Series Classification
- Collect more Review Papers
- I need to update the communication protocol document with the recent update emails

Questions that I have:

- Paper by Bagattini et al: I didn't understand the part of handling missing values from the timeseries
- Paper by Baydogan et al: I did't understand how did they discretize the CPE and created the histograms
- For now, do I have to read the details about complexity and understand the details of the algorithm ? or is it enough to understand how it works ?

20th July 2020, Communication Protocol and Literature Overview Document

Dear Prof Spiliopoulou,
I am sending an update for the week 13 jul - 20 jul: For the sake of easier communication, I created a project on overleaf where all my weekly work and updated documents will be. This way the latest version of everything I do will be there and we will not have to maintain different copies of documents sent via email.
Please use this link for the project: https://www.overleaf.com/read/frpfdfwhgkgc
Updates for the week:

- ☑ Datasets: Added information and summary about candidate datasets that I found along with a table describing each
- ☑ Libraries: Added the names of the 3 candidate libraries and a table describing the algorithms they implement
- ☑ Literature Overview: Got the papers of Ulf Leser and Bagattini et al, started reading them and adding summary points (haven't finished reading yet)
- ☑ Notes: Added some notes on distance measures for time series like Frechet, Edit distance and DTW. Also included old notes about stationarity of time series data
- ☑ TSCA Review: Included my summarization for the TS Bakeoff review paper.

Actions till next week:

- Finish reading the papers of Ulf Leser and Bagattini et al
- Check literature for correlated time series it might have some points useful to our problem

- Collect broader literature regarding Multivariate Time Series Classification
- Collect more Review Papers

Best Regards, Ismail Wahba

9th July 2020, Communication Protocol and Literature Overview Document

Dear Ismail Wahba.
Last part of my feedback before we meet in person, it concerns mode of communication and the literature overview document.

- Mode of communication:
  - I suggest that we exchange two documents:
    - your thesis draft (please pick the template from the KMD website)
    - a docx protocoll document (can be also latex, of course), where you write what we discuss at each date (do not forget to write the date), the points you wrote below for example. Then, my answers come and you incorporate them. And when an issue is closed, you mark it in gray. Questions from you also go to this document.
  - Some students use overleaf for their thesis document and googledoc for the protocol document. Depends on how much you trust google. You can consider alternatives.
- The literature overview document contains two parts - the concepts and the papers.
  - The concepts are fine, you have understood the terminology well.
    - ☐ I suggest that you create a separate document, sort of list-of-terms with explanations. It will become an appendix of your thesis.
  - The literature overview is a very good start, one paper per area. This will become part of the related work chapter.
    - ☐ So, I suggest that you start incorporating it into the thesis document.
  - Some of the papers are a bit old, there is a surge of publications in the recent years. Especially for early classification in TS: the 2009 paper is very old, there is a 2020 publication and two at least after 2018. So, keep these papers as basis, since you read and understood them.

☑ but now go to the literature overviews and in scholar.google and similar fora and collect also newer literature.

– You need a concept for doing literature research, eg what keywords you use, how you exclude papers, for which papers you follow citations, for which old papers you look for new citations to them; in the thesis, you must write down how you did that.

☑ There is an instrument called PRISMA. Please consider it.

Best regards, Myra Spiliopoulou

7th July 2020, Follow up on Literature Overview and preparing for first meeting

Dear Ismail Wahba.
We can indeed shift our meeting to earlier, to 15:00, same date 13. July (Monday). I am replying on some of the issues below, a more extensive mail comes later.

- As for the term "static data" I included in my notes, I actually myself did not really understand what it means in TS context and though about asking about it when we have our meeting to clarify what it means. From my readings, I understood that in order to predict future values for TS we need to "stationarize" data but I wasn't sure if this is related or not.

  ☑ Oh, it does not have to do with "stationary". Much simpler: when we study individuals (patients for example), we have their dynamic data, eg body temperature, blood sugar levels etc, recorded at each time point. These data constitute the multivariate time series. But we have also static data, eg the patient's birthdate. For the patients in the applications we investigate, there are questionnaires with several questions that are answered only once, since they do not change. Individuals can be deemed similar on the basis of these static data.

- I added the papers' citations in the overview document under each title (attached to this email)
  I have found 2 recent review papers, one of them is by Bagnall et al (2017) and the other is by students (2020) in Indian Institute of Technology (BHU). Is it better to go with more recent or more cited?
  I found 2 libraries that implement TS classification (pyts and sktime)
  I found a website which is dedicated for TS classification (Prof. Bagnall's team). It has datasets of different characteristics, I wonder what are the things I should focus on when choosing the dataset ?

  ☑ That it contains TRUE time series and multivariate ones. Sounds odd, but many of the datasets in the collections are not time series, they are images. Under `https://www.cs.ucr.edu/%7Eeamonn/`

`time_series_data_2018/` you will find in the 2nd column the type of the dataset. Ignore all that is called "Image", "Simulated", "Spectro" or "Device". I recommend that you concentrate on the type "Sensor" and look for multivariate ones.

5th July 2020 Preparation for first meeting

Dear Prof Spiliopoulou,
Thank you so much for the encouraging reply, I am really happy that my beginning was good enough and I hope to keep it up this way till the end.
July 13th at 16:00 is fine for me and I have no problem with the time of the meeting whether it is made earlier or not.
I will also send a reminder on that day in the morning as you requested.
As for the term "static data" I included in my notes, I actually myself did not really understand what it means in TS context and though about asking about it when we have our meeting to clarify what it means. From my readings, I understood that in order to predict future values for TS we need to "stationarize" data but I wasn't sure if this is related or not.
I also had some updates and questions regarding the todos for our next meeting:

- ☑ I added the papers' citations in the overview document under each title (attached to this email)
- ☐ I have found 2 recent review papers, one of them is by Bagnall et al (2017) and the other is by students (2020) in Indian Institute of Technology (BHU).
  - ☐ Is it better to go with more recent or more cited?
  - Bagnall, Anthony, et al. "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances." Data Mining and Knowledge Discovery 31.3 (2017): 606-660
  - Gupta, Ashish, et al. "Approaches and Applications of Early Classification of Time Series: A Review." arXiv preprint arXiv:2005.02595 (2020).
- ☑ I found 2 libraries that implement TS classification (pyts and sktime)
- ☑ I found a website which is dedicated for TS classification (Prof. Bagnall's team). It has datasets of different characteristics, I wonder what are the things I should focus on when choosing the dataset ?

Thank you in advance and I hope I am on the right track for this milestone
Best Regards, Ismail Wahba

29th June 2020 Literature Overview Feeback

Dear Ismail Wahba. Very good start! You interpreted the problem correctly, and your choice of papers is to the point. The overview of terminology at the beginning is good; just the term "static data" I did not understand in that context.
Some todos till then:

- ☑ I need also the citations of the papers, to know from where they are: the research field is very active and very mature, there show up many papers every couple of months. You will need to go for the state-of-theart only.
- ☑ Add the citations to the papers
- ☑ Find a survey that compares algorithms, so that you can choose the state of the art
- ☑ Find libraries with installed ts classification algorithms (the one by Bagnall et al?), so that you can use the library instead of developing the algorithms per se
- ☑ Find a TIMESERIES dataset that you can use for tasks 2 and 4, and for which you know the ground truth. In the donor's dataset we do not.

All the best, Myra Spiliopoulou

24th June 2020 Literature Overview

Dear Prof. Myra Spiliopoulou,
Please find in the attachment my literature overview.
I have included in this document some notes regarding time series classification problem in general and summary points for some papers that I skimmed through.
This is not a comprehensive list of all the papers I found or the algorithms that were mentioned in literature, but I tried to tackle the ones that sounded most important from my quick reading.
I have to say that TSCA is a completely new topic for me. I spent most of the time trying to understand what type of problem usually it is and familiarize myself with the different techniques that have been used for it, rather than digging deep into the the algorithms and understanding them in details.
But I believe that now I have somehow an overview of it and hope that things will get more clear while exploring more literature and also understanding the problem we have at hand.
I hope that what I have included in the document is enough to carry out our first appointment.
I would appreciate if you can send me the possible appointment time(s), so that I can coordinate things with my boss at work if there will be a conflict of times.

I also believe that due to the COVID-19 situation, the meeting might be on an online platform, so this is my Skypeid (isma3il.samir) in that case. If there is another platform/application that I should download, please just communicate it with me.
Best Regards, Ismail Wahba

16^{th} June 2020, Topic Introduction Email

"Performance degradation of time series classification algorithms (TSCAs) as the length of the time series diminishes".

- ☐ Goal is to build a testbed that compares TSCAs on a small set of ts, while shrinking the ts length.
- ☐ The testbed should encompass:
  - ☐ An algorithm that shrinks the length of the input time series in a non-random manner.
    We have algos from master theses, on which you can build.
  - ☐ A time series management tool that incorporates the algorithm 1 and stores the ts in a database.
    There are ready testbeds for time series classification, so an existing tool may be extended.
  - ☐ A collection of TSCAs
  - ☐ An evaluation utility that assesses degradation of the TSCAs across two dimensions:
    - ☐ model quality, assuming skewed distribution in an n-class problem
    - ☐ ranking of the variables that contribute to class separation.
      variables:= segments, shapelets or whatever the TSCAs use
  - ☐ A baseline for the donor's data-set:
    - ☐ It uses (latency,amplitude)_{ij}, for i in {leftear, rightear} and for j in {wave1, wave3, wave5} to separate across n=3 classes.
- ☐ The core work is on tasks 4 and 5.
- ☐ The thesis also demands a literature overview of TCSAs for multivariate time series and the implementation of a choice of TCSAs in task 3.
- ☐ The first tasks are (i) collection of literature on multi-var time series segmentation and classification, (ii) time-plan and (iii) preparation of a thesis proposal. Once you are midway in task (i), we make our first appointment.
- ☐ The donor's data-set is not available yet, we hope that it will be there within a couple of weeks. But as you see, there are further tasks to perform, so there is plenty of work to start with.