A common task framework for testing and evaluation at the Space Domain Awareness Tools, Applications, and Processing Lab

Imène R. Goumiri, Luc Peterson, Ashley Cocciadiferro, Ryan Lee, Jason Bernstein
Lawrence Livermore National Laboratory
Maj Sean Allen
U.S. Space Force (SSC/SZG)

LLNL-PROC-868468

ABSTRACT

The Space Domain Awareness (SDA) Tools, Applications, and Processing (TAP) Lab, or SDA TAP Lab, is an initiative of the U.S. Space Force to efficiently and effectively transfer technology from industry, academia, and federally funded research and development centers (FFRDCs) to Space Force guardians or operators. Participants in the SDA TAP Lab develop software for tasks such as determining whether a rocket launch might pose a threat to a satellite on-orbit, predicting future rendezvous and proximity operations, and detecting pattern of life violations. Testing and evaluation of this software is critical to ensure that it operates as required and so that it can be benchmarked against other software solutions. Lawrence Livermore National Laboratory is performing testing and evaluation for the SDA TAP Lab and drawing from best practices across the software development, SDA, and machine learning and artificial intelligence communities to ensure the process is quantifiable, objective, rigorous, and spurs innovation. In this paper, we outline the general approach we have adopted for testing and evaluation, namely the Common Task Framework that has driven innovation in artificial intelligence and machine learning, and focus on a particular benchmarking problem we have developed on predicting conjunctions.

1. INTRODUCTION

The Space Domain Awareness (SDA) Tools, Applications, and Processing (TAP) Lab is an initiative of the U.S. Space Force to accelerate the development by industry of capabilities to assist Space Force operators and guardians in space battle management. Tasks of interest include determining if a rocket launch is a potential threat to a satellite on orbit, detecting maneuvers, processing uncorrelated tracks, predicting future rendezvous and proximity operations (RPOs), and identifying photometry changes. Participants in the SDA TAP Lab, and in particular its Apollo Accelerator program, come from government, industry, academia, and federally funded research and development centers (FFRDCs). The goal of a group participating in a three-month Apollo Accelerator cohort is to develop a software product that addresses a specific task identified by the SDA TAP Lab as being of importance. For example, a group might train a machine learning (ML) algorithm on historic photometry data and use that algorithm to detect photometry changes in near real time.

Lawrence Livermore National Laboratory (LLNL) is testing, evaluating, and benchmarking capabilities developed in the SDA TAP Lab, and curating data sets for this purpose. There are three main reasons for benchmarking these capabilities. First, benchmarking quantifies the performance of algorithms and indicates to an operator using the software how well the algorithm performs. Second, quantifying algorithmic performance on standardized data sets allows comparisons to be made between different proposed solutions. Third, benchmarking allows the state of the art (SOTA) to be tracked over time, which further drives innovation as groups seek to improve on the SOTA. Benchmarking further needs to be accomplished transparently, objectively, and rigorously in order for the process to be trusted and utilized.

It is important to note that this testing and evaluation is not meant to be used for operational acceptance. Instead, it serves as a preliminary assessment of performance that can be used to guide further research and development.

[1] identifies the common task framework (CTF) as an approach that has led to significant progress in the field of predictive modeling. [1] credits [3] with first identifying the CTF in the context of a DARPA program on speech recognition. According to [1], the CTF is characterized by a group working on a common task with a particular data set, and whose solutions are evaluated using clearly defined and objective metrics. This approach is common in machine learning and artificial intelligence. For example, the MNIST and CIFAR-10 data sets are standard for testing ML classifiers and have been used to push the SOTA for decades. To date, there is no analogous, standard data set or common task for pushing the SOTA for SDA, despite a rapidly increasing literature on ML for SDA. The SDA TAP Lab therefore seeks to establish a CTF for driving SDA innovation and providing Space Force operators with the best tools available. However, the approach does have limitations that we discuss at the end of this paper.

There have been previous efforts at SDA benchmarking. For example, the European Space Agency (ESA) has a Kelvins project that hosts competitions on various SDA tasks such as streak detection and pose estimation, and the MIT ARCLab recently held a competition to identify satellite pattern of life violations [4]. The SDA TAP Lab effort is unique in that it is focused on tasks supporting space battle management for the U.S. Space Force and can use government-owned data that is not publicly releasable. Benchmarking challenges are open to groups participating in the Apollo Accelerator, which promotes collaboration and working together.

In the remainder of this paper, we describe the common task framework (CTF) and its adaptation to SDA, curation of data sets, and benchmarking of algorithmic performance. An example involving conjunction prediction is provided to illustrate the approach. Limitations of the approach are discussed along with actions to mitigate these issues.

2. COMMON TASK FRAMEWORK

We now describe the Common Task Framework (CTF) from [1] and discuss how it has been adapted to the SDA TAP Lab. Section 6.1 of [1] decomposes the CTF into data, task, and evaluation components, which are summarized in Fig. 1. The data set is fixed in the CTF, available to all challenge participants, and separated into input and response variables. There is a well-defined task to be accomplished with this data, such as developing an algorithm for making a certain type of prediction. Algorithms developed with the data for the intended task are then evaluated with established metrics for prediction performance on hold-out data. Since the data and evaluation metrics are fixed over time, progress on the task can be tracked and objective comparisons made between proposed solutions.

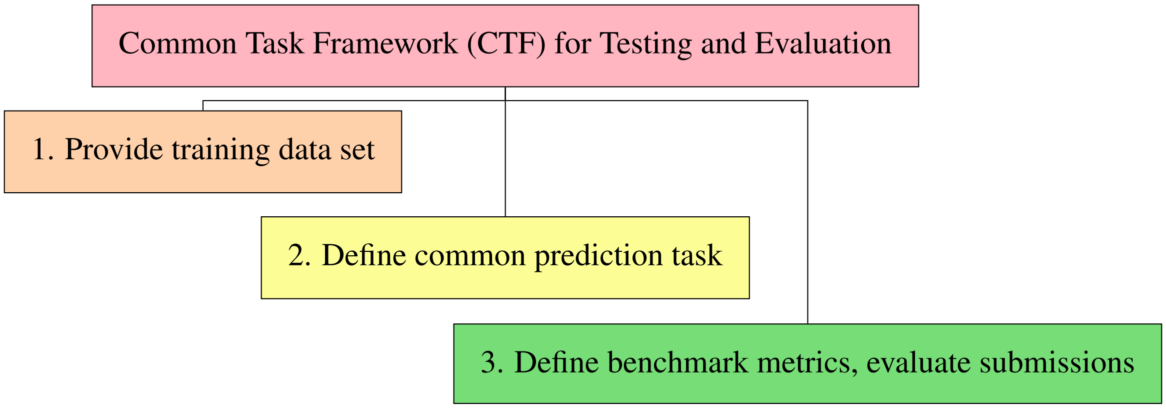

Common Task Framework (CTF) for Testing and Evaluation
1. Provide training data set
2. Define common prediction task
3. Define benchmark metrics, evaluate submissions

Fig. 1: The three aspects defining the common task framework as outlined in [1].

The CTF only requires slight specialization to be applied at the SDA TAP Lab. Tasks are obtained from Space Force operators who discuss their challenges with Apollo Accelerator participants. These challenges are translated into problem statements and common tasks that the participants work towards resolving. Data is obtained from the Unified Data Library (UDL) when relevant data for the task is available. When data is not available on the UDL, it is generated using modeling and simulation. The metrics used to score proposed solutions are standard, such as precision and recall, with some adaptation as discussed in sec. 4. Care is taken to ensure that the training and validation data sets have similar properties, such as the proportion of positive to negative cases.

The CTF provides a unified framework for improving on the SOTA for the different problems of interest to the SDA

TAP Lab. Without the CTF, groups could work on different data sets and with different metrics on the same task, making it impossible to compare their solutions and determine which is more suitable for further development. The CTF is so ingrained in ML research today through ML challenge platforms like Kaggle that it might seem like the only approach for coordinating effort across groups interested in making progress in a certain application area. However, this is not necessarily the case, as evidenced by the different conjunction prediction scenarios that can be found across the SDA literature. We emphasize that the CTF is central to our benchmarking approach in order to fairly compare methods and to identify and improve upon the state of the art.

3. DATA SET GENERATION AND CURATION

Identifying high-quality data sets for SDA benchmarking is the first step towards driving innovation in the common task framework. Our approach has been to use real-world data when possible and to use simulated data otherwise. For example, electro-optical measurement data from the Unified Data Library (UDL) was curated for a UCT processing challenge. The data provider is an industry participant in the SDA TAP Lab and was able to provide feedback and guidance on curation of the data set. In contrast, there is not a large amount of available data for predicting GEO direct ascent anti-satellite (ASAT) threats, and so simulation was used in this case. Multiple subject matter experts are involved at all stages of data set planning and curation to ensure that possible biases are identified and removed and that the algorithms that are being developed can take the data as an input. Each data set consists of multiple cases, ideally several hundred to thousand unique cases.

Following convention and best practice [2, ch. 7], data sets are divided into training, testing, and validation subsets. The complete training data set and solutions are provided to challenge participants to develop their algorithms and receive feedback on their performance. The validation data set provides the data for each case but not the solutions. Participants can submit predictions based on the validation data set and receive aggregate scores back, but not metrics for individual cases. The testing data set is withheld from the participants until the end of a competition. Participants can submit solutions for this data set a limited number of times and receive aggregate benchmark metrics back. The goal of the testing data set is to assess the performance of the algorithm on new data. The testing set therefore reduces the risk of over-fitting, since an algorithm that was tuned to score perfectly on the validation set through trial-and-error would likely do poorly on a new data set. Typical splits of a complete data set into training, validation, and testing subsets are 80%, 10%, and 10%, or 60%, 20%, and 20%.

4. BENCHMARKING METRICS

The metrics used to benchmark proposed solutions to the SDA challenges vary based on the specific task, but are often similar across tasks. For classification tasks such as identifying threats or rocket launches, standard binary classification metrics such as accuracy, precision, and recall are used. Regression tasks such as predicting the miss distance for an RPO are evaluated with metrics such as mean square error (MSE) and mean absolute deviation (MAD) error. These are standard machine learning metrics that are simple to explain and interpret. Other commonly-used metrics and loss functions such as binary cross entropy are also evaluated and can be used to rank methods, but are less directly interpretable.

Some cases in a particular data set can be more important to predict correctly than others for operators given SDA related tasks. In this situation, cases are weighted more in the metrics to reflect their increased importance. Consider an ASAT prediction task, for example. Having more warning time is beneficial to operators for this task, and so a weighting factor can be applied when computing the mean square error of the miss distance that down weights predictions for cases that have less warning time. A possible down weighting factor is one divided by the warning time, where the warning time is the time from the launch to the last observation. Hence, cases that have more warning time are down weighted in the overall score computed over all the cases. Binary classification metrics can be similarly down weighted by adjusting a constant multiplier for true positives or false negatives, say, based on warning time for a particular case.

5. CONJUNCTION PREDICTION CHALLENGE EXAMPLE

As a motivating example, we describe a common task framework approach to a benchmarking task involving predicting conjunctions between satellites. The SDA TAP Lab is focused on developing capabilities for closing several kill chains, including LEO and GEO direct ascent and co-orbital ASAT threats. Having the capability to predict conjunctions or close appraoches is a necessary component to close these kill chains. And while there is extensive literature on predicting conjunctions and established software and workflows for this task, a standard data set with real-world data collected from commercial sensors that motivates further development on conjunction prediction in threat scenarios is not readily available.

Fig. 2 summarizes the common task framework for the conjunction prediction challenge. The challenge is decomposed into training data, prediction task, and scoring components. This type of diagram quickly and succinctly describes benchmark problems of interest to the SDA TAP Lab.

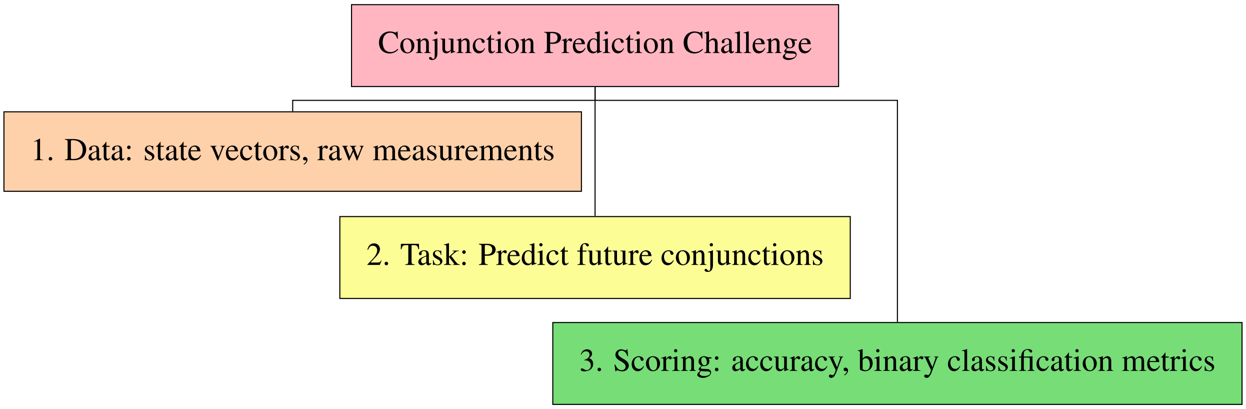
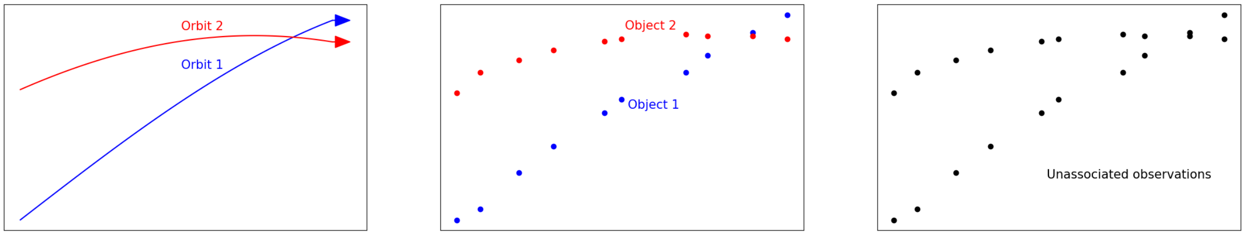

Conjunction Prediction Challenge
1. Data: state vectors, raw measurements
2. Task: Predict future conjunctions
3. Scoring: accuracy, binary classification metrics

Fig. 2: The common task framework for the conjunction prediction challenge.

The conjunction prediction challenge is actively being developed, but is planned to include easy, medium, and hard difficulty levels to encourage participants at all levels to benchmark and improve their capabilities. The challenges are differentiated as follows:

- **Easy Difficulty Level**: The common task is to predict conjunctions from a combination of two-line elements (TLEs) and state vectors, such as position and velocity vectors or Keplerian elements. To perform this task, participants need to propagate the orbits from the initial state vectors over a window of time and determine if a conjunction occurs over that time window according to a specified distance threshold such as 50 km. This challenge assesses orbit propagation fidelity and conjunction prediction accuracy.
- **Medium Difficulty Level**: The common task is to predict conjunctions from range or angle measurements on the objects, where each measurement is associated to one of the objects. In order to perform the conjunction assessment given this data, participants need to initially determine the orbits of the objects from the measurements and then propagate the orbits forward in time as in the easy conjunction challenge. Hence, this challenge assesses the quality of the orbit determination, orbit propagation, and conjunction prediction.
- **Hard Difficulty Level**: The common task is to predict conjunctions from range or angle measurements or tracks of multiple objects, where the data has not been associated to the objects. A catalog of orbits will be available, but it may not include all the orbits from which there are observations. Participants therefore need to associate tracks to objects and determine the orbits of the objects, propagate the orbits, and perform the conjunction assessment. This challenge assesses performance of data association, orbit determination, orbit propagation, and conjunction prediction.

Fig. 3 gives an idea of the data available for predicting a conjunction for each of the three difficulty levels. Note that the observations in Fig. 3b are contaminated by observation error, so that there will be error in the determined orbits. Also note that in Fig. 3c, some observations near where the orbits cross do not obviously belong to either of the orbits. These details make the conjunction risk assessment non-trivial and introduce variability in the performance of algorithms developed for the solution of this challenge. The important point is that the amount of information available for predicting a conjunction decreases with increasing difficulty level.



(a) Easy Difficulty Level (b) Medium Difficulty Level (c) Hard Difficulty Level

Fig. 3: Provided data for the conjunction benchmark challenge across difficulty levels.

In practice, the hard difficulty challenge is the most realistic, but our goal is to have a challenge be accessible to individuals with all levels of expertise. The performance metrics are also the same for each challenge, so starting with the easy or medium challenges are an effective way for participants to gain familiarity with the benchmarking app and metrics their submissions will be scored with. Note as well that the challenges build sequentially with the difficulty level, in that the medium challenge is the easy challenge with orbit determination, and the hard challenge is the medium challenge with data association. The challenges can thus be worked on in order as orbit determination and data association capabilities are developed.

The three conjunction data sets are made from a combination of real and simulated data, where the percentage of each data type varies by difficulty level. Conjunction data messages published on SpaceTrack.org are used to identify objects at risk of conjunction and predict conjunction times. For the easy challenge, publicly-available TLE data is readily available from websites such as SpaceTrack or Celestrak, and state vectors can be obtained from the Unified Data Library (UDL). However, radar or electro-optical measurements are not simple to obtain for pieces of debris identified in conjuction data messages, which complicates using real data for the medium and hard challenges. For these cases, we will take a modeling and simulation approach to producing synthetic range or angle data from the objects, using orbit propagation tools like the open-source SSAPy Python package [5]. This introduces errors into the benchmarking process because the simulation codes are not exact, which we account for by having conservative thresholds for identifying conjunctions. For each case, we plan on creating several hundred data sets where approximately half the data sets are of conjunctions, and half the data sets are not of conjunctions.

Several benchmarking metrics are computed for each challenge. Considering a data set of a conjunction to be a positive case and a data set without a conjunction to be a negative case, we compute the standard binary classification metrics including precision, recall, accuracy, and F1-score, which is the harmonic mean of the precision and recall. We also compute the root mean square error (RMSE) of the miss distance at the time of closest approach and the RMSE of the predicted time of closest approach. One issue is that warning time is important for operators that may have to make a decision to maneuver a satellite, for example, if a conjunction is predicted with high probability. We incorporate this consideration into the time of closest approach metric by adding a penalty for positive predictions that increases with decreasing warning time. Note as well that since there are multiple benchmark metrics that are computed for each submission, it is possible that one solution will not perform best in all metrics. This is a desired feature of the benchmarking approach in that it will reveal where different solutions perform best and expose precision-recall trade-offs, for example.

Conjunction prediction challenges and benchmarking are not new to the SDA community. The ESA hosted a similar challenge on their Kelvins platform in 2019 and the conjunction probability literature contains several benchmark problems and comparisons. The novelty here is that the training data is provided by industry and obtained through the UDL, and the intention is to develop capabilities that will ultimately be transitioned to the U.S. Space Force for space battle management. Utilizing the common task framework for this problem ensures that different approaches are objectively and transparently compared, and the best capability will be promoted on a path towards deployment to operators.

6. BENCHMARKING APP

A web application hosts benchmarking algorithms, provides access to the data sets and scores, and keeps track of submissions. This web application is intended to be hosted in a cloud environment controlled by the SDA TAP Lab and accessed by the TAP Lab participants. Users upload solutions for a particular data set in a specified csv format, which triggers benchmarking of the solutions. The benchmark metrics for the submission are stored in a database and displayed in a table in the web app that is visible to all users. Overall the app is similar to the Kaggle platform that hosts ML competitions, but is customized to the needs of the SDA TAP Lab and hosts SDA datasets.

The front end for the web app is built using NextJS, a React framework. The API is build using the ExpressJS framework, and the backend algorithms are built using Python. For any provided dataset, a leaderboard shows submissions ranked by various benchmarking metrics (e.g. accuracy, precision, and any custom metrics for that dataset). Benchmarking algorithms are built using Python and rely heavily on the Scikit-Learn and metrics modules. Users can either anonymously submit solutions or mark them with their name and organization name, which lowers the barrier to entry for participants that are in the early stages of developing solutions for a particular challenge.

7. LIMITATIONS

The common task framework is effective in part because it focuses research and development activity on a single problem over a long period of time. However, SDA challenges and problems change over time, which exposes a limitation of applying the CTF in this domain. For example, tracking objects in cislunar space is becomingly increasingly important for SDA, so a static challenge dedicated to tracking satellites in LEO or GEO could have an unintended consequence of directing resources away from this emerging challenge. Additionally, it is possible that a challenge can persist even after it has been effectively solved, with groups continuing to improve upon the SOTA performance even if the current SOTA method is sufficient for operational use. Operationally, there may not be a meaningful difference between object classification methods that have accuracies of 99.9% or 99.99%.

These limitations can be mitigated by periodically reviewing the challenges and assessing whether they should be updated or retired, and disincentivizing minor improvements against the SOTA. The Apollo Accelerator cohorts last three months, which provides a natural frequency for reviewing challenges and retiring them if they are no longer relevant or are effectively solved. To disincentivize over-fitting and improving the SOTA by an amount that is not operationally significant, we suggest retiring challenges that have been effectively solved with the hope that this encourages groups to pursue unsolved challenges. Another idea we are considering is regularly updating the training, validation, and testing data with current real world data, which guards against the issue of distribution shift in machine learning that can lead to the eventual failure of models trained on historic data.

8. CONCLUSION

A common task framework (CTF) has been adopted for testing, evaluation, and benchmarking at the U.S. Space Force's SDA TAP Lab. This framework is characterized by focusing sustained effort on optimizing clearly-defined performance metrics for a specified data set and task. Though this approach has arguably been implicitly used in the past, its formalization is necessary for the SDA TAP Lab since it takes place in the context of a large-scale, government-led technical accelerator involving industry, government, academia, and FFRDCs. Adaptation of the framework to the SDA TAP Lab is also necessary since the prediction tasks and considerations are distinct from those in other domains. Challenges of varying difficulty levels have been constructed for SDA TAP Lab participants with varying degrees of experience or expertise in SDA. Consideration has also been given to limitations of the approach, including distribution shift and diminishing returns as the state of the art model converges to the best possible model in terms of predictive performance. Initial work has been performed on establishing a benchmark data set and metrics for a conjunction prediction and uncorrelated track processing task, with future work planned on detecting rocket launches, performing cyber defensive operations for SDA, and reacquiring lost satellites.

REFERENCES

[1] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[3] Mark Liberman. Obituary: Fred jelinek. *Computational Linguistics*, 36(4):595–599, 2010.

[4] Peng Mun Siew, Haley E Solera, Thomas G Roberts, Daniel Jang, Victor Rodriguez-Fernandez, Jonathan P How, and Richard Linares. Ai ssa challenge problem: Satellite pattern-of-life characterization dataset and benchmark suite. In *Proceedings of the 24th Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS)*, 2023.

[5] T Yeager, K Pruett, and M Schneider. Long-term n-body stability in cislunar space. In *Proceedings of the Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*, page 208, 2023.

9. DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344.