Draft version February 22, 2023
Typeset using LaTeX default style in AASTeX631

Toward Classification of Fast Radio Burst Sources Using Unsupervised Machine Learning Techniques

Michael Gutierrez
California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125, USA

Project Mentor: Charles Steinhardt
Cosmic Dawn Center (DAWN)
Niels Bohr Institute, University of Copenhagen, Lyngbyvej 2, DK-2100 Copenhagen Ø

1. BACKGROUND

Since their first identification in 2007 (Lorimer et al. 2007), transient astronomical events known as Fast Radio Bursts (FRBs) have perplexed the scientific community. Believed to be extragalactic in origin, FRBs contain rich information about the parts of the universe they traveled through – possible clues to some of our biggest questions in both astronomy and cosmology. However, we have sparingly little clarity on their physical origins. At present, we can't even tell if they're all the same (as with Type 1A Supernovae, for example) or if there are several classes of FRB-emitting phenomena (as there are for merger events producing gravitational waves). FRBs lie at the frontier of our observational capabilities, making their study as challenging as it is interesting.

FRBs are, in fact, extremely fast (a few milliseconds at most); they occur in radio frequencies (RF, 500 MHz - 2 GHz); and in many cases, they come in unpredictable one-off bursts. There are estimated to be about 1000 detectable FRBs per day (Petroff et al. 2019), but considering how little of the radio sky is monitored at any given time, finding them is essentially luck-based. Even with dedicated detector arrays like the Canadian Hydrogen Intensity Mapping Experiment (CHIME), most FRB events lack enough corroborating observations to precisely characterize or localize their sources in the sky. For instance, while some FRB sources are known to periodically repeat, there is no known way of predicting if a burst came from a repeater using only the properties of the burst itself.[1] Additionally, the relevant L and UHF frequency bands are fraught with RF interference, which has led some to propose alternative, terrestrial explanations for FRBs detected at only one site (although many of these alternatives have been ruled out) (Petroff et al. 2015).

Despite these difficulties, more than 500 FRBs have been successfully catalogued by CHIME. Efforts to cross-reference them with data from other radio interferometry arrays (FAST, Jansky VLA, e-MERLIN, ASKAP, etc) as well as X-ray observatories (Chandra) have been mildly successful, and will continue to improve with more wide-field FRB discovery machines like the DSA-2000 (Hallinan et al. 2019) and GReX (Connor et al. 2021) spinning up. So far, a few tens of candidates have been identified as possible repeat events (Wang et al. 2022), and some were found to have coincided with X-ray bursts captured in the same part of the sky, allowing them to be traced to a host galaxy (Ravi et al. 2019) (or possibly, as Connor et al. (2016) showed, from within our own galaxy). These observations have spawned a multitude of theories about the phenomena responsible for FRBs. No conclusive evidence for any of them has been found, although as of today, some theories (young pulsars/other supernova remnants) appear more likely than others (distant alien spacecraft, white holes) (Petroff et al. 2019).

This is a common roadblock: a relative wealth of data, but few footholds for scientific insight. Especially in the case of projects like CHIME, the resulting catalogs of objects and events are too complex to offer any hope of human readability, so it is necessary to use computational tools to distill them down to the features we care about (Amiri et al. 2018). Unsupervised machine learning (ML) techniques such as Random Forest, Principal Component Analysis

magutier@caltech.edu

[1] There are a few interesting empirical theories which would be worth cross-checking with the CHIME catalog. See Caleb et al. (2023) and Pleunis et al. (2021).

(PCA), and time-distributed Stochastic Neighbor Embedding (t-SNE), which specialize in extracting patterns from unlabeled datasets, have proved effective in dimensionality reduction of photometry and spectroscopy data for the Sloan Digital Sky Survey (Baron & Poznanski 2016) and COSMOS catalogs (Hovis-Afflerbach et al. 2021). Most relevant for this project, however, are clustering techniques. These use the results of dimensionality reduction to visually arrange the input data on a plot with distances between points indicating their similarity (e.g. Fig. 1). When tuned correctly, these algorithms can reveal previously undetected correlations and trends (Reis et al. 2021), which makes them well-suited for extracting and separating a mix of data points with subtly different properties.

A notable application of clustering to astronomical data was a successful dimensionality reduction of Gamma Ray Burst (GRB) observations from NASA's *Swift* mission (Jespersen et al. 2020). GRBs are similar to FRBs in their brevity, difficulty of detection, and mystery; the fact that ML algorithms were able to clearly classify GRBs into short- and long-types (see Jespersen et al., Fig. 1) is incredibly promising. A similar attempt was made by Zhu-Ge et al. (2022) to classify FRB sources using clustering algorithms, supporting a distinction between repeating and non-repeating FRBs. However, while they did obtain successful clustering into about 5 to 7 groups, they were unable to work back from those results to explain the physical differences. This could suggest that overfitting was occurring, which indicates that the input data was overly simplified. In other words, more work is needed to verify the separation.

In this proposal, we outline an endeavor to classify FRB sources using unsupervised ML tools. With the release of CHIME's Catalog 1 (Amiri et al. 2021), there is an equal abundance of data available and unanswered questions. By building on the techniques used by Jespersen et al. and Zhu-Ge et al., we hope to uncover evidence of an observable and physically meaningful heterogeneity of FRB-emitting phenomena. Such a separation would better inform follow-up studies of burst events as well as future observation strategies. It is possible that the amount or quality of data in the catalog is not sufficient to evoke any statistically significant patterns with these methods, but if it is, the field would greatly benefit from them.

2. OBJECTIVES

The aim of this project is to determine if there are multiple classes of FRB-emitting phenomena. To accomplish this, we'll need to tease out any hidden patterns within available FRB observations which may have evaded human inspection – or to be able to confidently conclude that there are none. Given the success of unsupervised ML strategies in parsing high-dimensional astronomy data, we are hopeful that this, too, is a puzzle that can be cracked with proper care given to the methodology.

In particular, we want to minimize human and systematic biases in the information supplied to the ML algorithms. For some data sets, this might necessitate pre-processing the data to remove irrelevant features and emphasize certain connections over others, as Jespersen et al. and Zhu-Ge et al. did. In other cases, though, the best results are best obtained without modifying the data at all. We suspect that the latter applies to the CHIME catalog, given that the summary data on their website consists mostly of derived quantities. As the CHIME collaboration explains in the Catalog 1 publication (Amiri et al. 2021), the main difficulty in working with this data is accounting for the selection biases of the detector array itself. In light of this, our first roadblock to overcome will be developing a proper pre-processing workflow. It will need to be rigorous enough to maintain the natural structure of the data set and, ideally, flexible enough to be applicable to other similar catalogs released in the future.

Hand-in-hand with data pre-processing is the consideration of which ML tools to use. It is important to note that with large inputs, these programs take a non-trivial amount of time to run, so we will have to be selective with which combinations of pre-processing methods and ML "hyperparameters" to try. For example, the t-SNE algorithm takes a number of arguments with the most important being *perplexity*, which essentially allows us to assign relative weights to small/local scale variations vs. large/global scale trends (see Fig. 1). We can use this in tandem with knowledge of the aforementioned selection biases and existing hypotheses about FRB sources to narrow down the set of subspaces within the data set which could contain patterns of interest. With any luck, this process will eventually reveal a clear separation of FRB categories *and* be able to identify the empirical evidence on which it is based, enabling a reassessment of the theory from a new perspective.

3. APPROACH

The majority of the computation involved will be performed and visualized using Python (especially the **scipy** and **scikit.learn** (Pedregosa et al. 2011) packages) as well as the Uniform Manifold Approximation and Projection

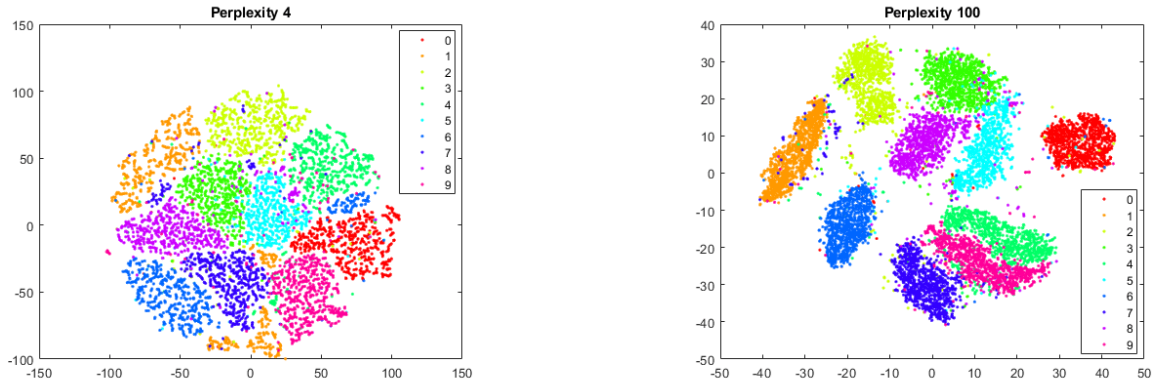

Figure 1. Example of the effects of the *perplexity* hyperparameter on the output of the t-SNE algorithm run on the MNIST database. The x- and y-axes do not represent any meaningful physical units or properties; they are essentially a parameter space. Image source: https://www.mathworks.com/help/stats/tsne-settings.html

(**UMAP**) algorithm (McInnes et al. 2018). The bulk of our time on this project will be spent interacting with these programs, so we'll take the opportunity to build proficiency with them and gain intuition for how to tune the "hyperparameters" to data sets appropriately. Figure 1 shows an example of how a sub-optimal choice of the *perplexity* parameter affects the quality of the results. (In this case, the figures *are* the results of the algorithm; if they were given to us without the color-coding, we would have a much better chance of differentiating two separate classes with the Perplexity 100 clustering.) This is largely a trial-and-error process, but it will nonetheless require careful assessment of the outputs at each step. If there is no clear separation of classes, or if the separation turns out to be based on irrelevant features, then recalibration of the model will be necessary.

Even with those strategies, the algorithms may still fail to find a separation depending on how the input data was formatted. In fact, we would expect this from a data set containing only summary properties such as the one Zhu-Ge et al. used from https://www.chime-frb.ca/catalog. The quantities in that table were derived from the raw recordings of the FRBs (which are also available). While the data in this format is more human-readable, a large amount of the information is thrown away in the process. To that end, we plan to take advantage of the full data set from the CHIME catalog. However, we can't simply feed in *everything* – each one of the > 500 cataloged events consists of over a gigabyte of real-time recorded radio frequencies and localization estimates (Amiri et al. 2021, sec. 3.2). We will have to evaluate which subsets of the full catalog are the most relevant and least noisy through a combination of studying the CHIME publication and inspecting the data visually. A particular 'dimension' will be useful to us if (1) the values it takes on are causally connected to the FRB emitter; (2) it is invariant to terrestrially- or systematically-induced variations; and (3) it is available in the same format for all data points. For example, the estimated sky location of a burst is likely irrelevant to its physical nature[2]. On the other hand, the radio frequency data is vital to identifying the properties of the burst, but it also contains a large amount of noise. In the case of Jespersen et al., it turned out that each GRB signal had an essentially random time offset from zero. In order to filter out this influence, a Fourier transform was applied to the entire set before attempting any classification. We will likely have to devise similar strategies to clean up our data.

Finally, as mentioned, we will need to implement an efficient way to visualize the outputs of the algorithms to make them human-readable at a glance. This will take the form of using Python plotting libraries to label the resulting clustering graphs with the known properties of each point. In a successful separation of emitter classes, we would expect to see a differentiation between repeating sources and non-repeating sources, for example. We would not expect to see repeated emissions from the same source classified into different categories.[3] Using this strategy of verifying emergent patterns with current theory, we can direct the ML algorithm slowly but surely into a useful classification of FRB sources.

[2] If considered alongside the burst's distance measure, it could indicate that the source resides in a particular galaxy.

[3] Probably.

4. WORK PLAN

- **Before the summer:** Stay up-to-date with literature on FRB observations and analyses in order to hit the ground running
- **Weeks 1-3:** Gain proficiency with tools of the trade (Python/Jupyter, scikit-learn, UMAP) as well as the CHIME FRB catalog
- **Weeks 4-9:** Develop a standard renormalization scheme for burst data; attempt dimensionality reduction and clustering using a variety of embedding strategies; modify algorithm settings and hyperparameters and repeat as necessary
- **Weeks 10-11:** Identify the most successful results; inspect any clusterings for common trends/characteristics; as appropriate, hypothesize physical interpretations

REFERENCES

Amiri, M., Bandura, K., Berger, P., et al. 2018, The Astrophysical Journal, 863, 48, doi: 10.3847/1538-4357/aad188

Amiri, M., Andersen, B. C., Bandura, K., et al. 2021, The Astrophysical Journal Supplement Series, 257, 59, doi: 10.3847/1538-4365/ac33ab

Baron, D., & Poznanski, D. 2016, Monthly Notices of the Royal Astronomical Society, 465, 4530, doi: 10.1093/mnras/stw3021

Caleb, M., Driessen, L. N., Gordon, A. C., et al. 2023, Discovery of an as-yet non-repeating fast radio burst with the hallmarks of a repeater, arXiv, doi: 10.48550/ARXIV.2302.09754

Connor, L., Sievers, J., & Pen, U.-L. 2016, Monthly Notices of the Royal Astronomical Society: Letters, 458, L19, doi: 10.1093/mnrasl/slv124

Connor, L., Shila, K., Kulkarni, S., et al. 2021, Publications of the Astronomical Society of the Pacific, 133, 075001, doi: 10.1088/1538-3873/ac0bcc

Hallinan, G., Ravi, V., Weinreb, S., et al. 2019, The DSA-2000 – A Radio Survey Camera, arXiv, doi: 10.48550/ARXIV.1907.07648

Hovis-Afflerbach, B., Steinhardt, C. L., Masters, D., & Salvato, M. 2021, The Astrophysical Journal, 908, 148, doi: 10.3847/1538-4357/abd329

Jespersen, C. K., Severin, J. B., Steinhardt, C. L., et al. 2020, The Astrophysical Journal, 896, L20, doi: 10.3847/2041-8213/ab964d

Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J., & Crawford, F. 2007, Science, 318, 777, doi: 10.1126/science.1147532

McInnes, L., Healy, J., & Melville, J. 2018, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv, doi: 10.48550/ARXIV.1802.03426

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825. http://jmlr.org/papers/v12/pedregosa11a.html

Petroff, E., Hessels, J., & Lorimer, D. 2019, The Astronomy and Astrophysics Review, 27, doi: 10.1007/s00159-019-0116-6

Petroff, E., Keane, E. F., Barr, E. D., et al. 2015, Monthly Notices of the Royal Astronomical Society, 451, 3933, doi: 10.1093/mnras/stv1242

Pleunis, Z., Good, D. C., Kaspi, V. M., et al. 2021, The Astrophysical Journal, 923, 1, doi: 10.3847/1538-4357/ac33ac

Ravi, V., Catha, M., D'addario, L., et al. 2019, Nature, 572, 352, doi: 10.1038/s41586-019-1389-7

Reis, I., Rotman, M., Poznanski, D., Prochaska, J., & Wolf, L. 2021, Astronomy and Computing, 34, 100437, doi: 10.1016/j.ascom.2020.100437

Wang, F., Zhang, G., Dai, Z., & Cheng, K. 2022, Nature Communications, 13, doi: 10.1038/s41467-022-31923-y

Zhu-Ge, J.-M., Luo, J.-W., & Zhang, B. 2022, Monthly Notices of the Royal Astronomical Society, 519, 1823, doi: 10.1093/mnras/stac3599