RATE GRB-z: Ranking High-z GRB Candidates using Random Forest Classification on Early-Time Metrics

A. N. Morgan*, J. Long, T. Broderick, J. W. Richards, J. S. Bloom

University of California, Berkeley - Center for Time Domain Informatics†

*amorgan@berkeley.edu

†http://cftd.info

Abstract

As the number of observed Gamma-ray Bursts (GRBs) continues to grow, follow-up resources need to be used more efficiently in order to maximize science output from available telescope time. As such, it is becoming increasingly important to be able to rapidly identify bursts of interest using early-time metrics. Here we present our Random forests Automated Triage Estimator for GRB redshifts (RATE GRB-z) used for rapid identification of high-redshift (defined here as z > 4) candidates using early-time metrics from the three telescopes onboard the *Swift* satellite. This classifier will provide a recommendation - based on available telescope time - of whether a new burst should be observed. Our training set consists of 136 Swift bursts with known redshifts, only 17 of which are z > 4: an imbalance which presents several statistical challenges. Cross-validated performance metrics on the training data suggest that ~50% of high-z bursts can be captured from following up the top 20% of our ranked candidates. We further applied the method to 200+ Swift bursts with no known redshift to rank order the bursts according to their likelihood to be high-z.

Introduction

In principle, indications of high redshift are present in quickly available metrics from the three telescopes onboard Swift (BAT, XRT, UVOT) such as detections of the afterglow in the optical, and others. While past studies have used hard cuts on a small number of these attributes with some success (e.g. [1,2,3]), we aim to improve upon these techniques by utilizing machine learning algorithms.

We collated data on all Swift GRBs up to and including GRB 100621A directly from GCN notices and automated pipelines ([4,5]) that process and refine the data into more useful metrics. Tens of attributes were parsed from the various sources and collated into a common format. Short bursts, and bursts without rapid notices from the BAT, XRT, and UVOT were removed from the sample for uniformity. This leaves 348 events: 136 with known redshift, and 17 with z > 4.

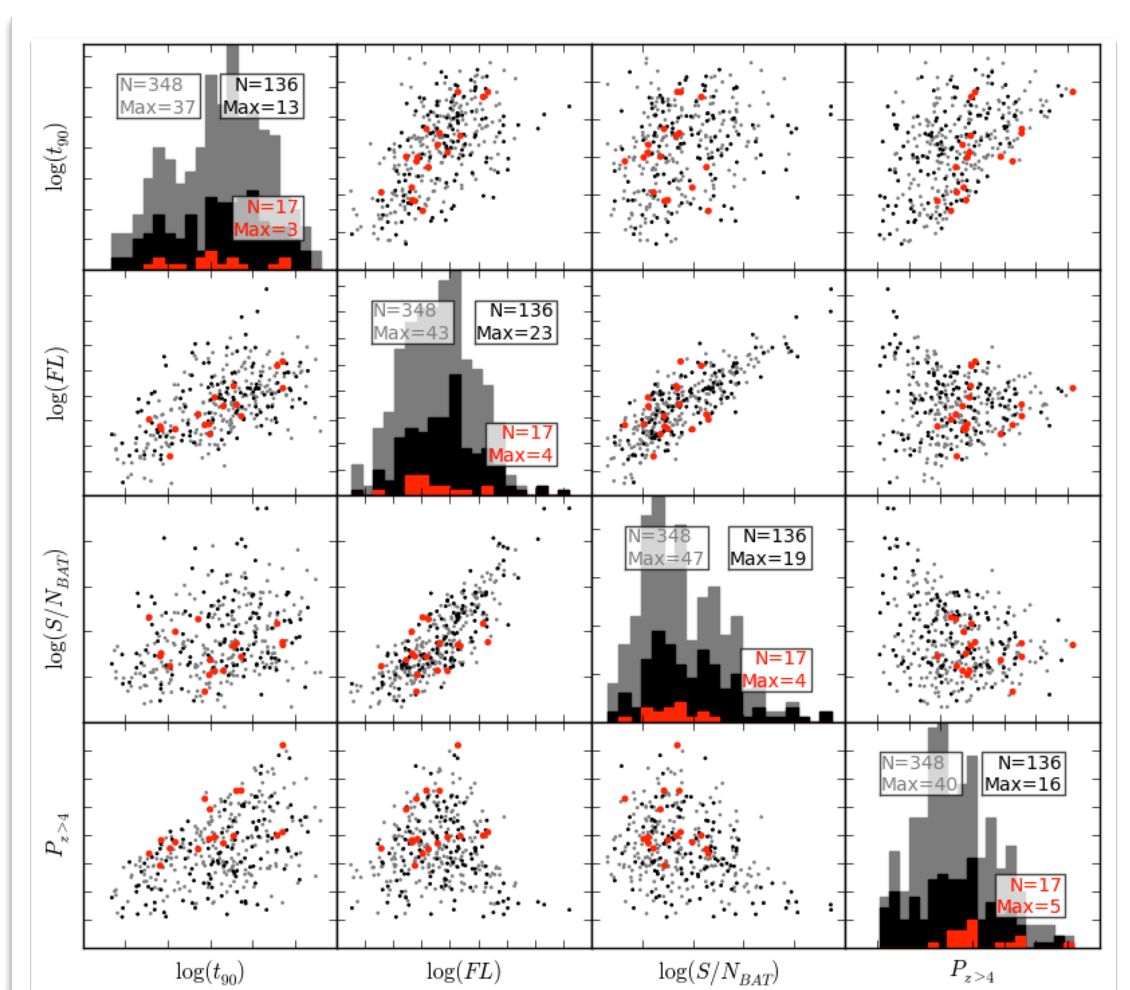

Figure 1: A selection of early time high-energy features derived from BAT data. The grey points show the full distribution of *Swift* GRBs. Bursts with known redshifts are black, and events with redshifts greater than 4 are over-plotted in red. A total of 12 features were used in our final classifier.

Methodology

We use Random Forest (RF) for our classifier due to its ability to select important features, resist overfitting the data, model nonlinear relationships, and handle categorical variables [6]. RF is an ensemble classifier that averages together the results from many iterations of Classification and Regression Trees (CART). A CART tree is constructed by splitting observations into two groups based on values of a given feature. RF then averages together many CART trees, each time running the CART algorithm on a bootstrap sample of the data, considering only a subset of the features at each split.

Our primary goal is a decision for each new GRB: should we devote further telescope observing time to this burst or not?

The RATE GRB-z method is as follows: Let Q be the fraction of bursts one has telescopic resources to follow up on. We rank the GRBs in the *training* set by probability of being high-z using ten-fold cross validation [7]. We obtain a probability of high-z for *new* events by inserting them into the RF classifier. Let Q' equal the fraction of bursts in the training set that have a higher probability of being high than this new burst. With proper calibration (Fig. 2), this leads to a simple decision point:

If Q' ≤ Q, we recommend follow-up for the new burst.

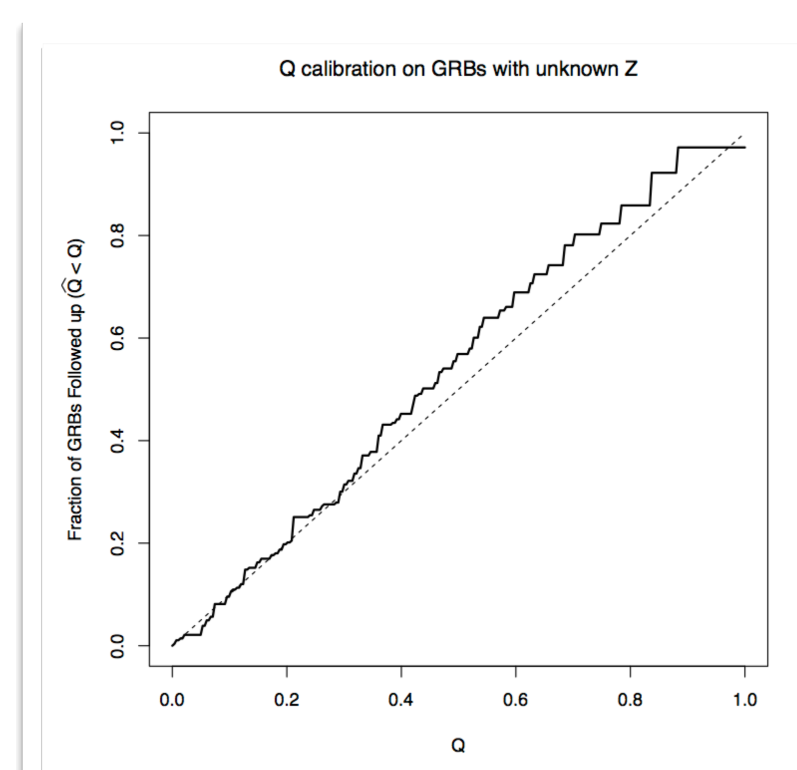

Figure 2: Calibration of Q: how well the user-desired follow up fraction Q correspond to the actual number of bursts recommended to be followed up by the algorithm (Q' ≤ Q). This plot shows the calibration on the test set of bursts without a known redshift (212 total events). Despite coming from a different population of GRBs, Q' closely traces the expected Q value.

Results

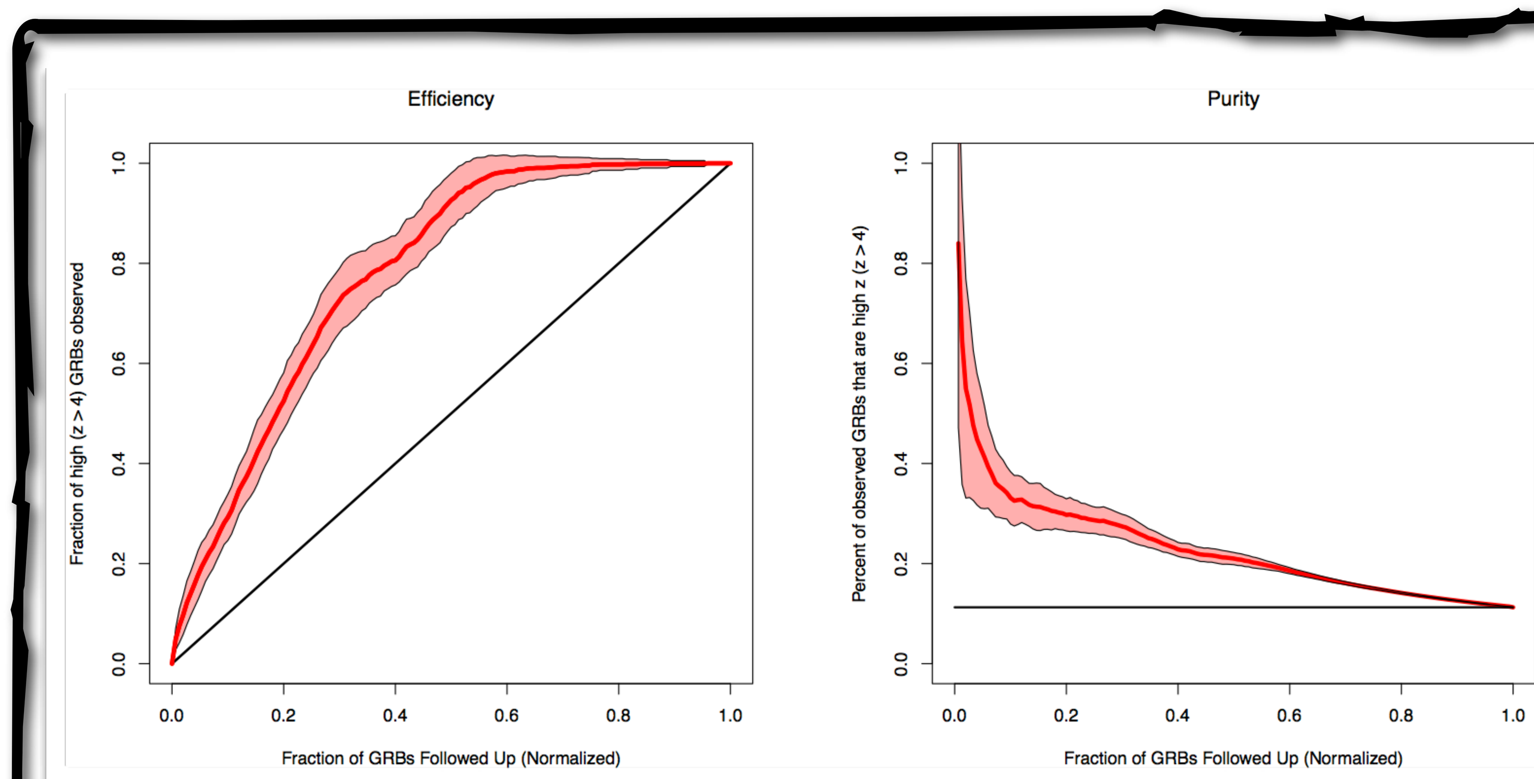

Figure 3: Cross-validated performance metrics on our training data showing efficiency ($N_{\text{high observed}}/N_{\text{total high}}$; left panel) and purity ($N_{\text{high observed}}/N_{\text{total observed}}$; right panel) vs fraction of GRBs followed up according to our decision criterion (Q) with a high-z cutoff of z=4. Seventeen bursts (~12.5% of our training set) are z > 4.0. The curve uncertainties shown are 1 standard deviation from the mean value averaged over 100 RF seeds.

Several data challenges were encountered and accommodated for:

Small dataset: (136 training GRBs): Controlled by minimizing useless feature usage (which hurt classifier performance).

Class imbalance: (12.5% high-z) Controlled by punishing misclassifications of high-z events more strongly through class-weighting.

Missing features: Controlled with imputation on missing values.

Figure 3 shows the cross-validated performance metrics for our final classifier, which suggests that with the RATE GRB-z method:

By following up on the top 20% (Q' < 0.2) of new GRBs, one can capture ~50% of all high-z (>4) events. Also, we expect ~30% of these followed-up GRBs will be high-z.

To facilitate the dissemination of high-z predictions to the community, we have set up a website (http://swift.qmorgan.com) with Q' rankings for past GRBs, and an RSS feed (http://swift.qmorgan.com/rss.xml) to provide real-time results from our classifier on new events.

References: [1] Grupe, D. et al., 2007, *AJ*, 133, 2216; [2] vanden Berk, D. E. et al., 2008, *AIPC*, 1000, 80; [3] Ukwatta, T. N. et al., 2009, *AIPC*, 1133, 437; [4] Butler, N. R. & Kocevski, D. 2007, *ApJ*, 668, 400; [5] Butler, N. R. et al., 2007, *ApJ*, 671, 656; [6] Breiman, L. 2001, *Machine Learning*, 45, 5; [7] Kohavi et al. 1995 *IJCAI*, 2, 1137

Sponsored by an NSF-CDI grant " (award #0941742) "Real-time Classification of Massive Time-series Data Streams" (PI: Bloom).