Searching for Gravitational Waves in LIGO Data: Using Tools in Multivariate Analysis

Ariana Sage Minot

National Science Foundation,

International REU Program Hosted by University of Florida

Cardiff University, Wales

Duke University,

Trinity College of Arts and Sciences

(Dated: August 5, 2008)

Evidence of gravitational waves associated with gamma ray bursts(GRBs), in addition to perhaps revealing new physics, may explain the cause of these highly energetic events. In order to detect the presence of gravitational waves in association with GRB events, it is necessary to separate signal and background effectively. Different multivariate analysis algorithms using Boosted Decision Trees(BDTs) and Artificial Neural Networks(ANNs) were implemented to achieve better separation during this project. Using simulated events of gravitational wave signal and real LIGO data not associated with GRBs, trees and networks were trained and tested under many different configurations. Currently, the best classifier is a BDT that achieves a signal efficiency of $89.5 \pm 0.2\%$ at 1% background contamination.

I. DETECTING GRAVITATIONAL WAVES IN LIGO DATA ASSOCIATED WITH GRB EVENTS AND THE CURRENT NATURE OF OUR SEARCH

The goal of this search is to analyze the presence of gravitational waves that concur spatially and temporally with GRB events recorded by electromagnetic satellites.

A. What is a Gamma Ray Burst?

A GRB is an intense flash of γ rays, which are waves of the highest energy and frequency within the electromagnetic spectrum. These events are observed to be isotropically distributed and are characterized by length of duration. Short GRBs last up to 2 seconds, and long GRBs last more than 2 seconds. GRBs are believed to be the most highly energetic events that have occurred in the universe since the Big Bang, however the progenitors of GRBs are not well known. See Figure 1. Observations show that long duration GRBs are associated with star forming galaxies, and theories suggest they are caused by supernovae events. The engines, or causes, of short duration GRBs are not as well understood. Some theories suggest that the merger of binary systems composed of either two neutron stars(NS/NS) or a neutron star and a black hole(NS/BH) produces a short duration GRB.

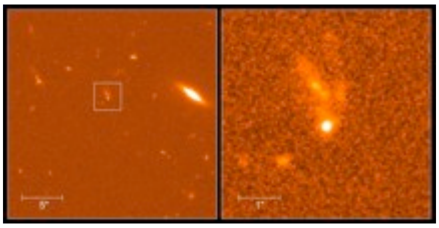

FIG. 1: Optical afterglow of gamma ray burst GRB-990123 (the bright dot within the white square and in the enlarged cutout) on 23 January 1999. This image was taken with the NASA Hubble Space Telescope Imaging Spectrograph (STIS).

B. Why should GRBs be associated with gravitational waves?

Sources of GRBs are also thought to be sources of gravitational waves. Some leading theories suggest that during and around the GRB event, gravitational waves are being emitted. Data whose GPS time matches up with the times of known GRBs is analyzed for the presence of gravitational waves. The form of the gravitational wave should carry a signature of where it came from, so physicists hope that evidence of gravitational waves at these times will provide insight into what caused these GRBs.

C. Organizing and Segmenting the Data

There are 213 GRBs recorded by satellite with GPS times accurate to within a second during LIGO's fifth science run, dating Novemeber 2005 to October 2007. The data is divided into segments called off source and on source data. See Figure 2. On the timeline below, the star represents the GPS time of the GRB as recorded by electromagnetic satellites. An asymmetric interval of three minutes is placed around the GRB event, and this data is referred to as on source data and will be searched for gravitational wave signal. On either side of the on source data is what is termed the off source data so that the entire interval comprises three hours of data. There are different models for possible sources of GRBs, and each model predicts different times that gravitational waves will be emitted in relation to the GRB event. The on source data is asymmetrically distributed, because some models predict gravitational wave emission in the period leading up to the GRB, such as the merger of a binary neutron star system. With respect to the current theories of gravitational wave emission, this three minute interval is conservative. The off source data is used for background analysis.

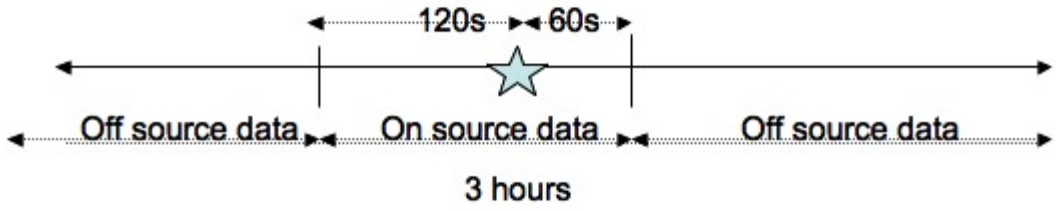

FIG. 2: Details of how the data is partitioned for analysis into off-source and on-source data.

D. Event Generation and Noise Analysis

The first step in the analysis is to generate events by looking for transient fluctuations in the data that may be due to a gravitational wave. To generate events, search code is run on both the off source and the on source data and returns events that pass certain criteria for being a gravitational wave. Next, the search code is rerun on the data after adding injections, or simulated gravitational waves, to see how well the code performs at finding the gravitational waves. A waveform called a sin-gaussian with energies within a certain range is used for injections since the exact form of the gravitational wave is not known.

To assess the noise present in the data, the off source data is analyzed under the assumption that the noise polluting the off source data will look the same as the noise polluting the on source data. The off source data is broken into three minute segments. Consistency tests, which are a series of vetoes, categorize the events as being either signal or background. See Figure 3. These vetoes currently are designed more or less by hand. To design the veto, two dimensional plots are made of several different measures of energy that characterize each event. Using events known to be background from the off source data and events known to be signal from the injections, various plots of one energy as a function of another can be made, and the veto is determined by seeing how the signal and background distributions behave in relation to these variables.

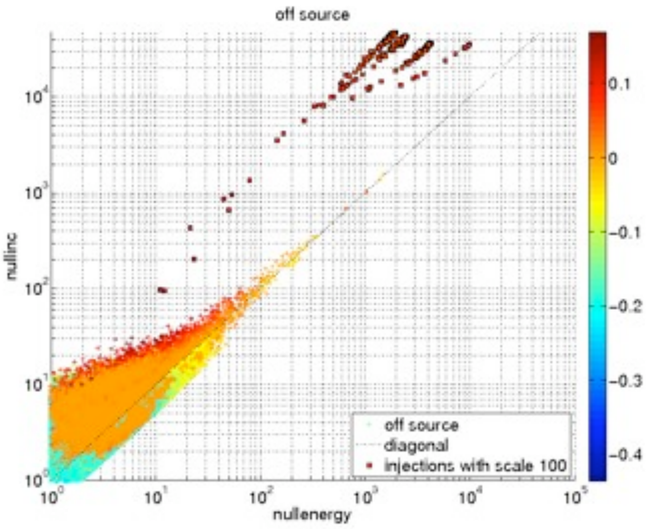

FIG. 3: This image was made using the X-Pipeline codes of the current analysis. This is a plot of null-energy versus null-inc using off-source data events and simulated injection events. Null-energy and null-inc refer to two different energies that characterize the events. The background and signal distributions have distinct relationships with respect to these two variables, in particular that the background events have a high correlation between null-energy and null-inc. Relationships such as these are used to design the series of vetoes in the consistency tests.

Finally, the 'loudest' event in each three minute segment of off source data is recorded. The 'loudest' event is the background event that looks the most like signal by having a large value of energies for both polarizations of a gravitational wave and having a high correspondence in both detectors.

II. INCORPORATING MULTIVARIATE ANALYSIS INTO THE CURRENT ANALYSIS PIPELINE

The current search can be improved by letting the computer automate large portions of the anaylsis. Instead of hand tuning vetoes, consistency tests are replaced with a classifier, which is some form of an artificial intelligence that looks at several input variables and classifies events as either signal or background. The loudest event is determined by using the classifier output value, a number typically between 0 and 1 that represents the classifiers confidence that an event is either signal or background. A classifier ouput value of 1, for instance, indicates that the classifier believes with a very high amount of confidence that the event is signal. The classifier also has the added advantage that it incorporates all information about the events, not just the relationships between variables determined by two-dimensional plots.

A. Steps of a Multivariate Analysis

There are two main stages to the multivariate analysis conducted in this project, training and application. The training step is responsible for the initial configuration of the classifier. The classifier is given a training sample, made up of injections and off-source data, where each event is known to be either signal or background. The signal was composed of a sample of injections of strengths in proportion to what is expected from real data. For example, injections of high strength are less common than injections of low strength. During the training phase, the classifier performance is evaluated by considering how often the classifier mistakes background for signal and how efficiently the classifier correctly marks signal events. To optimize classifier performance, different configuration options are available. Choosing the options that best configure the classifier is very dependent on the application at hand, and much of the effort of this project was put into testing different configurations.

III. OVERVIEW OF THE MULTIVARIATE TOOLS USED

A. Boosted Decision Trees(BDTs)

Performance of Boosted Decision Tree classifiers has been the most effective so far. In a BDT, events are placed in a root node. See Figure 4. Then, the algorithm considers each variable that characterizes the event. The variable that provides the best separation is used to perform a veto, and the algorithm also determines the value of the variable that will provide the best separation. The process is repeated at each node, and the algorithm may choose a different variable to perform the veto at different nodes. The nodes continue to split until they contain a certain number of events specified by the user, and a node is classified as signal or background by whichever it contains a majority of.

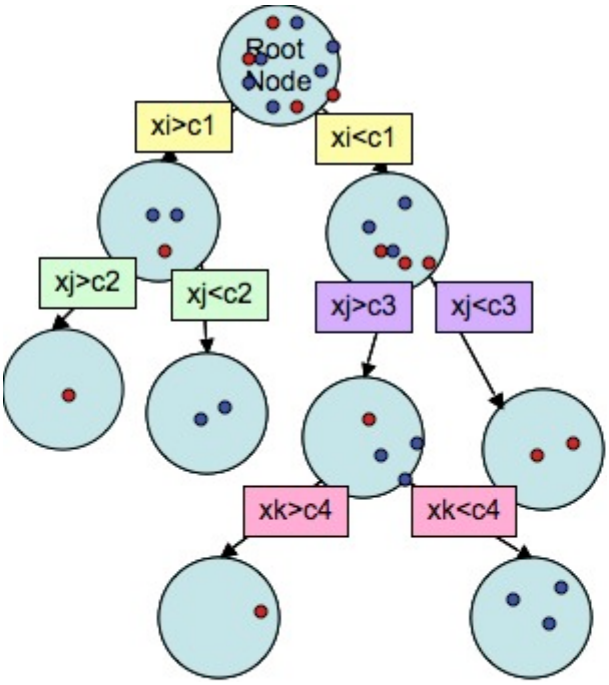

FIG. 4: Schematic representation of a Boosted Decision Tree inspired by the TMVA Users' Manual. The events, represented by blue and red dots, are placed in the root node. At each level of the tree, the variable with the best separating power is used. This example shows a highly effective tree since all of the red dots and blue dots are separated from one another by the end of tree. In practice, there will be many more blue and red dots and some contamination of one color in an end node with majority of the other color dot.

BDTs are also highly susceptible to 'overtraining'. If the classifier becomes too sensitive to statistical fluctuations in the training data, it will not perform as well on an independent testing sample. To compensate for this shortcoming, a forest of decision trees is constructed, and each event is classified by taking the majority vote of the classification done by the trees in the forest. The trees in a forest are all derived from the same training sample, and events undergo a process called boosting, where the weighting of the events is modified to increase performance of the classifier and statistical stability. A boosted classifier is one that is trained repeatedly using a reweighted, or boosted, training event sample each time. All of the individual classifiers configured during the training session are combined to form the final classifier. The boosting method used in this project is called AdaBoost, adaptive boosting. AdaBoost gives a higher weight in subsequent trainings to events that were misclassified during the training of the previous tree. The best classifiers thus far have used a Boosted Decision Tree algorithm.

B. Artificial Neural Networks(ANN)

Artificial Neural Networks were also explored as a classifying tool during this project. An ANN is a computational or mathematical model based on biological neural networks. ANNs consist of a set of artificial neurons that model complex relationships among variables to find patterns in the data. The output of each neuron returns a sigmoid function of the input variables. See Figure 5. A sigmoid is a non-linear S shaped function, with a range of values between 0 and 1.

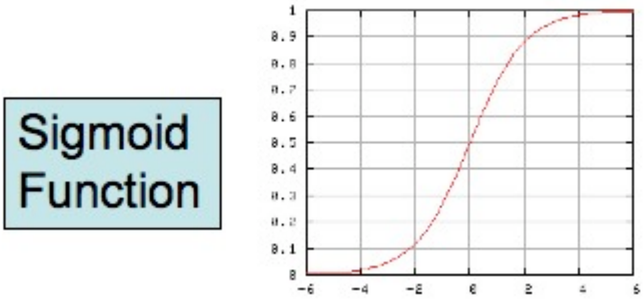

FIG. 5: Output of each individual neuron.

ANNs are usually made of several hidden layers, and neurons only communicated with adjacent layers. See Figure 6. The input of each neuron is a weighted sum of the outputs of the previous neurons, and training a neural networks involves adjusting the weighting of the output signals to obtain the best separation of background and signal events.

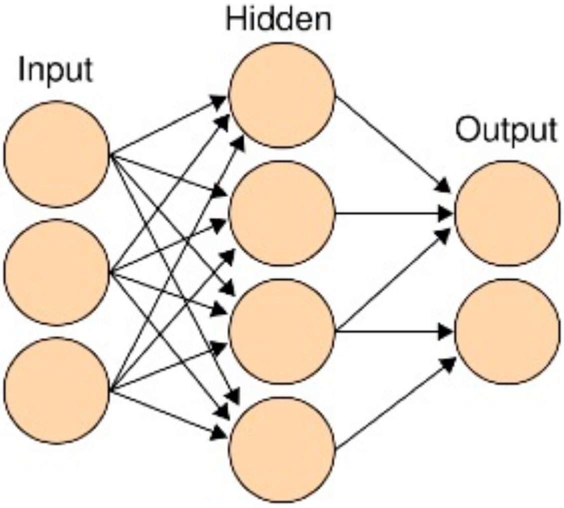

FIG. 6: Schematic representation of an Artifical Neural Network inspired by TMVA Users' Manual.

IV. RESULTS

Here, we consider the first results of applying multivariate classifiers to offsource data and simulated gravitational waves.

A. The Physical System Represented by the Data Used in this Analysis

The physical system being considered in this analysis is that of GRB 070201, which is a short duration GRB whose sky position is coincident with the spiral arms of the Andromeda galaxy, M31. During this GRB, the two detectors at the LIGO Hanford site were in use and collecting data. See Figure 7. The data of the GRB was February 1, 2007 at a GPS time of 854378604.

B. Efficiency and Plot for the Currently 'Best' Configured Boosted Decision Tree

Many different configurations of classifiers were tested during this project. To decide how well a classifier performs, the efficiency of signal for 1% background contamination and the

FIG. 7: Interferometer detectors at LIGO Hanford Site in Washington.

amount of overtraining were taken into account. See Figure 8. The plot below is of the classifier output value for a signal and background sample. These signal and background samples were applied to the trained classifier that had the best performance, i.e. most favorable signal efficiencies and least amount of overtraining. There is a vast number of different configurations available so it remains to be seen if this classifier has truly optimal performance. At this point in the project, it is the best performing classifier. The signal efficiency for a 1% background contamination is approximately 89%.

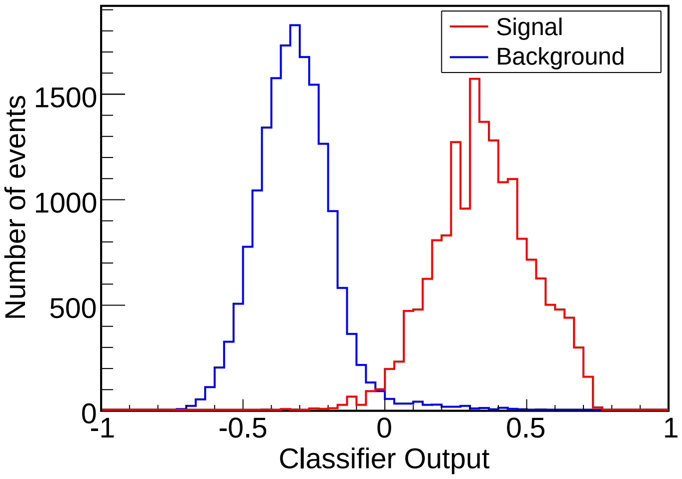

FIG. 8: The classifier ouput represents how confident the classifier is that an event is signal. Higher classifier output values correspond to signal events. The blue distribution is background, and the red distribution is signal. 99% of the background events lie to the left of a classifier output value of approximately 0.2, and $89.5 \pm 0.2\%$ of signal events lie to the right of 0.2. Therefore, the signal efficiency at 1% background contamination is $89.5 \pm 0.2\%$.

C. Can Multivariate Analysis be applied to the current search?

In order to integrate use of the classifier into the current analysis pipeline, the loudest event is described as the background event with the highest classifier output value. A set of loudest background events is accumulated by taking the loudest event in each three minute interval of off-source data. Next, a cumulative distribution is made of the fractions of 'loudest' background events versus classifier output value. See Figure 9.

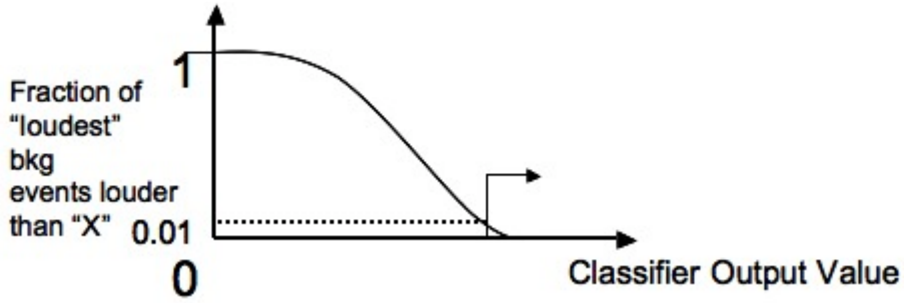

FIG. 9: A cumulative distribution of the fraction of loudest background events louder than some classifier output value, "x", versus classifier output value. Looking at this plot, we can say, for example, that if there is an on-source data event with a classifier output value lying to the right of the arrow, there is a 1% chance that a background event would produce an event this loud.

D. Making an Upper Limit on Gravitational Wave Amplitude

If there are no events of a certain loudness in the on source data, it can be said with a certain amount of confidence that there was no gravitational wave stronger than some amplitude or it would have been found. For example, this analysis can be used to make a statement, such as, "A gravitational wave of strength greater than some amplitude, h, would have produced an on-source event with a clasifier output larger than some value say 90 % of the time. Therefore, we exclude or limit the gravitational wave emission to be less than h at 90 % confidence level." No detection of gravitational waves is still physically interesting, because it indicates that either the model of how the gamma ray burst is produced is incorrect or that the gamma ray burst is actually further away than previously thought.

V. CONCLUSION

It is hoped that the integration of multivariate analysis classifiers into the current data analysis pipeline will improve the sensitivity of the search for gravitational waves in LIGO data associated with gamma ray bursts.

[1] http://en.wikipedia.org/wiki/Gamma_ray

[2] http://en.wikipedia.org/wiki/Image:Gammarayburst-GRB990123.jpeg

[3] Implications for the Origin of GRB 070201 from LIGO Observations, LIGO Scientific Collob-oration, K.Hurley,2007, arXiv:0711.1163v2

[4] http://grwavsf.roma1.infn.it/PSS/antennas/Ligo_LA2.jpg

[5] TMVA User's Manual. http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf

[6] Using Boosted Decision Trees to Separate Signal and Background in $B \rightarrow Xs$ Decays, James Barber, SLAC Program.