Machine Learning in Gravitational Wave Analysis

J. Andrew Casey-Clyde

May 9, 2017

Abstract

Gravitational wave detectors such as LIGO and Virgo offer new insights into our universe, yet also present challenges for analysis, primarily due to the sensitivity of the detectors to terrestrial sources of movement, as well as to the volume of data generated by these detectors. However, given the natures of these challenges, they are well suited to machine learning solutions. In this paper, I present techinques for the identification of transient events in LIGO timeseries data on all channels, as well as techniques for classifying these events as either terrestrial or non-terrestrial in origin.

1 Introduction (rework needed)

In 1916, based on his theory of general relativity, Einstein first predicted the existence of gravitational waves [1]. 100 years later, in 2016, the first two gravitaitonal wave events were confirmed using data from Advanced LIGO [2] [3], finally confirming the century old prediction. However, while it was the data from Advanced LIGO which led to this confirmation, this is not the first gravitational wave data to come from LIGO (and in fact, at the time of this writing, the Advanced LIGO data is not publically available). And while Advanced LIGO does indeed boast improved sensitivity as compared to its previous iterations, this does not preclude the possibility that there may be gravitational wave events hidden in previous data releases (and at the very least, these techniques would also be applicable to Advanced LIGO data, once it is released). Considering the less sensitive nature of older LIGO data, and it's impact on searching for gravitational waves, as well as the fact that these transients are fairly short lived, it makes sense to apply machine learning techniques to attempt to find these signals, instead of trying to search these datasets by eye, which is both, far more error-prone and much more time consuming.

In fact, multiple machine learning techniques have been explored specifically for the purposes of processing LIGO data, including both Artificial Neural Networks (ANN) [4] and Difference Boosting Neural Networks (DBNN) [5]. These techniques can be used, in conjunction with auxiliary channel data, to identify non-astrophysical transients, which may then be removed from the dataset. Once these datasets are removed, the remaining transients can then be further classified (such as by using hierarchical clustering algorithms) in order to determine potential transient categories, as well as how related those categories are morphologically.

2 Methods

In order to classify transients in the LIGO data, they must first be found. To this end, I apply the Kleine Welle algorithm, which decomposes wavelets into the time-scale domain and searches for regions of energy overdensities in the signal [4][6]. These techniques may be applied to all detector channels to extract transients, and in fact will be necessary for differentiating between terrestrial and astrophysical transient sources.

2.1 Data Cleaning

The LIGO gravitational wave data is characterized by Gaussian noise, with non-Gaussian transient events. As such, before applying the Kleine Welle algorithm to search for overdensities, it is useful to first whiten the data by taking it's Fourier transform and dividing by the amplitude spectral density (ASD), which is the square root of the power spectral density, and then transforming back [7]. This helps to suppress signal due to noise, enhancing transient signals. An example of the pre and post whitened ASD for the segment of data around GW150914 is shown in Figure 1.

Additionally, a butterworth bandpass filter may be applied to the whitened data [6][7] in order to remove regions of large or fluctuating background noise, such as those regions outside $80 - 300$Hz in Figure 1a.

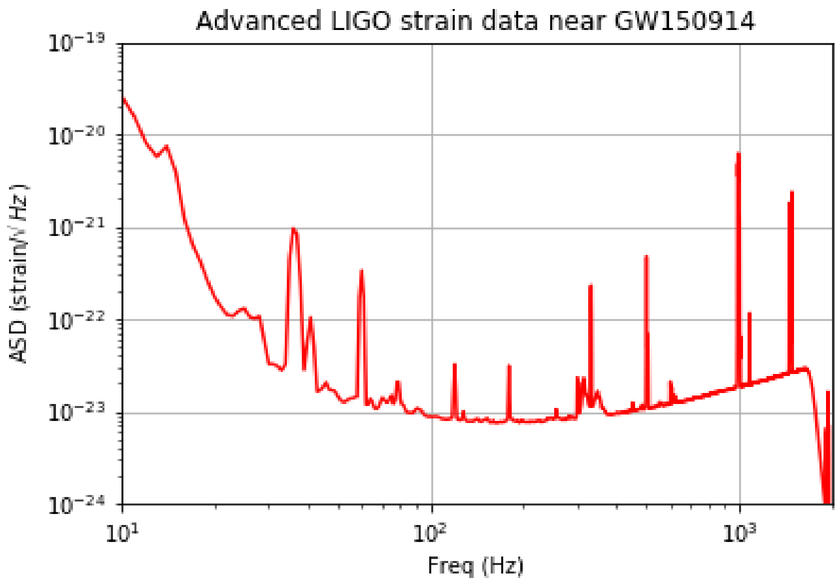
2.2 Event Identification

Once the data has been whitened and bandpassed, it can be fed through the Kleine Welle algorithm. The algorithm works through the application of a dyadic wavelet decomposition, whose coefficients can be used to search for signal energy overdensities [4][6]. In general, the wavelet transform is defined by the equation [6][8]

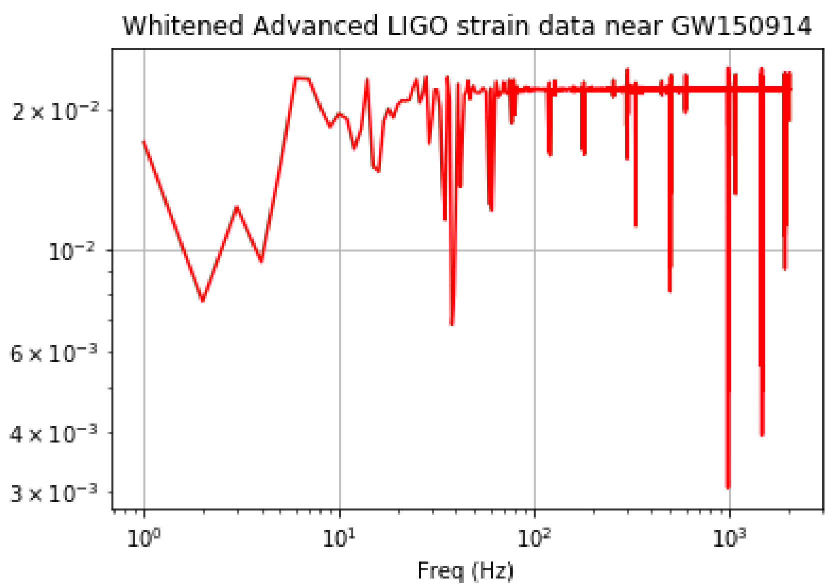
$$W_f(u,s) = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt, \tag{1}$$

where s is the scale. The calculation of this transformation can be made computationally inexpensive through the discretization of s to a dyadic sequence, such that $s \in \{2^j\}_{j \in \mathbb{Z}}$ [6][8].

Applying this to the LIGO timeseries data at multiple scales j [9] yields a series of detail coefficients D_j for each scale, with elements d_{ij}, whose values scale with signal energy and approach a zero mean Gaussian [6]. We can threshold on these elements to look for outliers by using the standard deviation of each decomposition level, σ_j, such that $d_{ij}/4\sigma_j > 4$. Elements which pass this



(a) ASD before whitening



(b) ASD after whitening

Figure 1: ASD vs. frequency for the Advanced LIGO data near GW150914, before and after whitening. Notice that in about the 80 − 300Hz range, the noise is at a minumum, and relatively constant. This is the range we wish to search for gravitational waves.

threshold are then normalized to create a set of square normalized coefficients such that

$$E_j = D_j^2/\sigma_j^2 = \{\epsilon_{ij}\}, \tag{2}$$

where the elements ϵ_{ij} correspond to the same elements d_{ij}. Taking only those ϵ_{ij} whose corresponding detail coefficients passed the earlier thresholding, we can cluster these in the time-scale domain (after rescaling both axes to the same unit size). The mean shift clustering method is particularly well suited to this problem, as it does not require prior knowledge or the number of clusters present [10], which in principle we cannot know for a given segment. An example of this clustering is shown in Figure 2. The elements of a given cluster C can then be added to form the total normalized cluster energy [6],

$$E_c = \sum_{(i,j)\in C} \epsilon_{ij}. \tag{3}$$

The significance of a cluster is then defined as [6]

$$S = -\ln \int_{E_c}^{+\infty} \chi_N^2(E)dE, \tag{4}$$

where the number of degrees of freedom N is the number of elements in the cluster. With this, it simply becomes a matter or thresholding the significance of each cluster.

2.3 Event Classification

Once transients have been identified in the strain data, we can use events that are correlated with hardware injections to train our classifier. I am currently deciding on a classifier to use, which is a bit behind where I would like to be. I was planning on using an ordered veto list in combination with auxiliary channel strain data to differentiate between terrestrial and non-terrestrial transients, and had done research towards this, only to find today that while the public LIGO data releases include some auxiliary channel data, it does not contain enough for these purposes. It does contain information about times where hardware injection was active, which can be used to train a neural network for event identification using these modeled events, and is the direction I plan on going.

3 Analysis

Keeping in mind that each LIGO science run contains data collected using progressively more sensitive instruments, I choose to apply these techniques to the S6 data release [11], from the most recent data collection run before Advanced LIGO instruments came online, which offers the best sensitivity for publically available data.

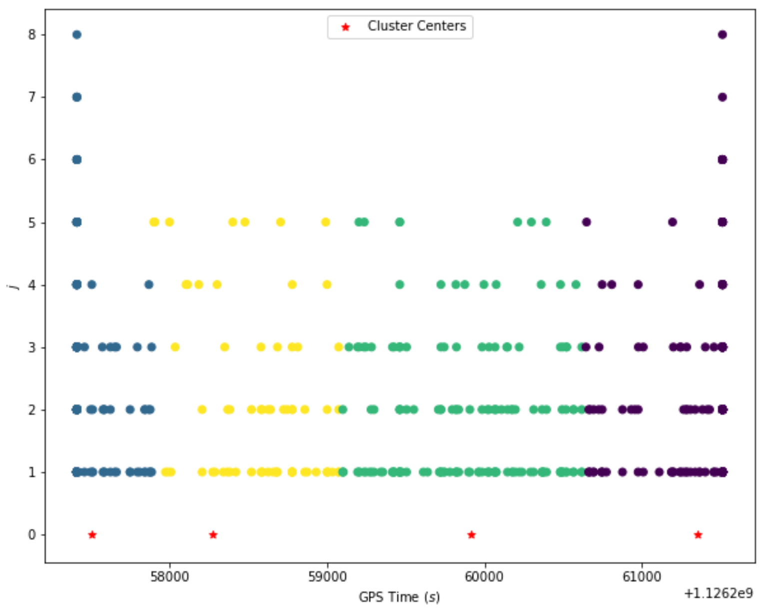


Figure 2: Square normalized coefficient clustering in the timeseries region around GW150914

References

[1] Albert Einstein. Näherungsweise Integration der Feldgleichungen der Gravitation. *Sitzungsberichte der Königlich Preußischen Akad. der Wissenschaften (Berlin), Seite 688-696.*, 1916.

[2] The LIGO Scientific Collaboration and Others. GW150914: First results from the search for binary black hole coalescence with Advanced LIGO. feb 2016.

[3] B.P. Abbott and Others. GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary Black Hole Coalescence. *Phys. Rev. Lett.*, 116(24):241103, jun 2016.

[4] Rahul Biswas, Lindy Blackburn, Junwei Cao, Reed Essick, Kari Alison Hodge, Erotokritos Katsavounidis, Kyungmin Kim, Young-Min Kim, Eric-Olivier Le Bigot, Chang-Hwan Lee, John J. Oh, Sang Hoon Oh, Edwin J. Son, Ruslan Vaulin, Xiaoge Wang, and Tao Ye. Application of machine learning algorithms to the study of noise artifacts in gravitational-wave data. *Phys. Rev. D, vol. 88, Issue 6, id. 062003*, 88(6), mar 2013.

[5] Nikhil Mukund, Sheelu Abraham, Shivaraj Kandhasamy, Sanjit Mitra, and Ninan Sajeeth Philip. Transient Classification in LIGO data using Difference Boosting Neural Network. sep 2016.

[6] Lindy Blackburn. KleineWelle Technical Document, 2007.

[7] LIGO Scientific Collaboration and Jonah Kanner. SIGNAL PROCESSING WITH GW150914 OPEN DATA.

[8] Stéphane Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 2nd edition, 1999.

[9] Filip Wasilewski. PyWavelets - Wavelet Transforms in Python PyWavelets Documentation.

[10] Željko Ivezić, Andrew J. Connolly, Jacob T. VanderPlas, and Alexander Gray. *Statistic, Data Mining, and Machine Learning in Astronomy.* Princeton University Press, Princeton, New Jersey, 2014.

[11] LIGO Scientific Collaboration. LIGO Open Science Center release of S6, 2015.