EPJ Web of Conferences **245**, 06021 (2020) https://doi.org/10.1051/epjconf/202024506021
CHEP 2019

Utilizing Unsupervised Machine Learning in BSM Physics Searches At The LHC

Adam Leinweber[1,*] and Martin White[1,**]

[1]University of Adelaide, North Terrace, Adelaide, SA 5005

Abstract. Recent searches for supersymmetric particles at the Large Hadron Collider have been unsuccessful in detecting any BSM physics. This is partially because the exact masses of supersymmetric particles are not known, and as such, searching for them is very difficult. The method broadly used in searching for new physics requires one to optimise on the signal being searched for, potentially suppressing sensitivity to new physics which may actually be present that does not resemble the chosen signal. The problem with this approach is that, in order to detect something with this method, one must already know what to look for. I will showcase one machine-learning technique that can be used to define a "signal-agnostic" search. This is a search that does not make any assumptions about the signal being searched for, allowing it to detect a signal in a more general way. This method is applied to simulated BSM physics data and the results are explored.

1 An Introduction to a Standard LHC Analysis

Recent searches for beyond the standard model (BSM) physics at the ATLAS and CMS experiments of the LHC [1] have been unsuccessful. In a standard LHC analysis, one must begin with a signal definition which is then optimised on, removing as much of the standard model background as possible to uncover the signal. This method has a high efficiency for signals that resemble the model chosen for optimisation, but requires one to already know what to look for. If nature has not chosen the same standard model extension as the signal definition, the search will not find anything. A basic outline of the steps involved in this method are as follows.

1. Select a signal and final state.
2. Model all relevant background processes (for example, by Monte Carlo methods).
3. Optimise kinematic selections on functions of the four momenta in the given final state, to define regions of the data that have a high signal-to-background ratio for the simulated signal.
4. Compare a detailed background estimate in the signal regions with the observed yield in the LHC data, and determine the statistical significance of any noted excess.

*e-mail: adam.leinweber@adelaide.edu.au
**e-mail: martin.white@adelaide.edu.au

2 Performing an Analysis Utilizing Unsupervised Machine Learning

We instead propose using a "signal agnostic" search method in which one merely looks for non-standard-model-like processes instead of searching for an assumed signal model. This is possible to do using unsupervised machine learning [3], namely anomaly detection. The process that we propose is similar to the method detailed prior, but with a few key changes. Note that it is necessary to apply a minimal preselection to ensure that selected events are compatible with detector trigger requirements. It is also arguably easier to search each final state separately in order to uncover anomalies, to make the background estimation easier to obtain in each search. This does introduce some model dependence, but this preselection is kept intentionally minimal so as to introduce as little model dependence as possible.

1. Model background processes, creating training and testing datasets.
2. Apply a minimal preselection.
3. Train an unsupervised anomaly detection algorithm on the simulated background and obtain an "anomaly score" calculator.
4. Calculate the anomaly scores for both the simulated background and observed LHC data.
5. Compare the event yield in the high anomaly score regions of the simulated background with the observed yield in the LHC data, and determine the statistical significance of any noted excess.

We have compared the performance of a number of unsupervised machine learning techniques, using Monte Carlo simulations of supersymmetric signals and their dominant background processes. In the presentation at CHEP 2019, we showed results for two hypothetical signal processes, one containing a 404 GeV Stop quark process and one containing a 1 TeV Gluino process [4].

The variables that the anomaly detection algorithm is trained on are the p_T and ϕ of the missing energy, the p_T, η, and ϕ of each particle and jet, as well as physical variables m_T and H_T. p_T, η, and ϕ are the transverse momentum, pseudorapidity and azimuthal angle of a given particle. m_T, or the transverse mass, is defined as $m_T = \sqrt{2p_T E_T^{miss}(1 - \cos(\Delta\phi))}$, where p_T is the transverse momentum of the most energetic lepton, E_T^{miss} is the missing energy in the transverse plane, and $\Delta\phi$ is the angle between the lepton and E_T^{miss} in the transverse plane. H_T is defined as the scalar sum of jet p_T. Particles and jets not present in a given event are zero padded.

2.1 The Isolation Forest

The anomaly detection algorithm that was shown at CHEP 2019 is called an isolation forest. First outlined in Ref [5], the isolation forest works by creating trees which recursively divide the dataset in an unsupervised way. Imagine a dataset $P = \{P_1, ..., P_n\}$ of n events with d variables, such that P_i can be expressed as $P_i = (x_1, ..., x_d)_i$. P is divided into two subsets by randomly selecting an index q such that $1 \leq q \leq d$ and a value p such that $min(x_q) \leq p \leq max(x_q)$, where $min(x_q)$ and $max(x_q)$ are the minimum and maximum values of x_q across the entirety of P. Events with $x_q < p$ are placed in a dataset P_l and events with $x_q \geq p$ are placed in a dataset P_r. The dataset is recursively divided in this way until either: (i) the tree reaches

a height limit, (ii) $|P| = 1$ or (iii) all events in P have the same value. The anomaly score of a given event is calculated as

$$s(P_i, n) = 2^{-\frac{E(h(P_i))}{c(n)}}, \tag{1}$$

where $h(P_i)$ is the tree depth of an event P_i, $E(h(P_i))$ is the average of $h(P_i)$ from a collection of trees, and the normalisation factor $c(n)$ is given by

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right), \tag{2}$$

where $H(i)$ is the harmonic number as estimated by $\log(i) + 0.5772156649...$ (The Euler-Mascheroni constant). The logic is that a highly anomalous event should take fewer slices to isolate than a totally typical event, and thus have a higher anomaly score. The isolation forest is particularly useful because of its linear time complexity and low memory requirements. Other algorithms have been tested and the isolation forest was found to have the best performance.

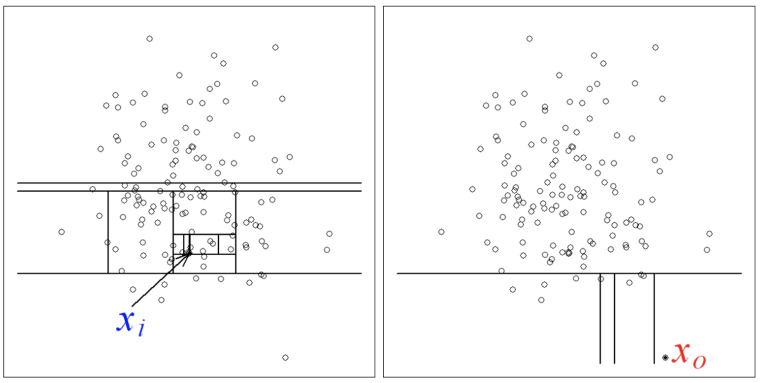

Figure 1. A visual representation of the logic of the isolation forest algorithm. The blue point x_i is quite typical of the dataset and thus requires more slices to isolate, compared to the red point x_o which is more anomalous and thus requires fewer slices to isolate.

3 Results

Results for two different signal models are presented here. The two signals are a 1 TeV gluino signal, and a 404 GeV supersymmetric top (stop) quark signal. The gluino signal was chosen as it is very easy to differentiate from the standard model background. This makes it a good demonstration and preliminary check of the method. The stop quark signal is a very difficult signal to find as it is kinematically very similar to top pair production.

3.1 Gluino Results

The gluino results, shown in Figure 2, demonstrate that, as expected, there are almost no signal events in the low anomaly score region of the histogram where the background dominates. In the high anomaly region there are many more signal events but unfortunately the event counts are not high enough to discover a signal in this region.

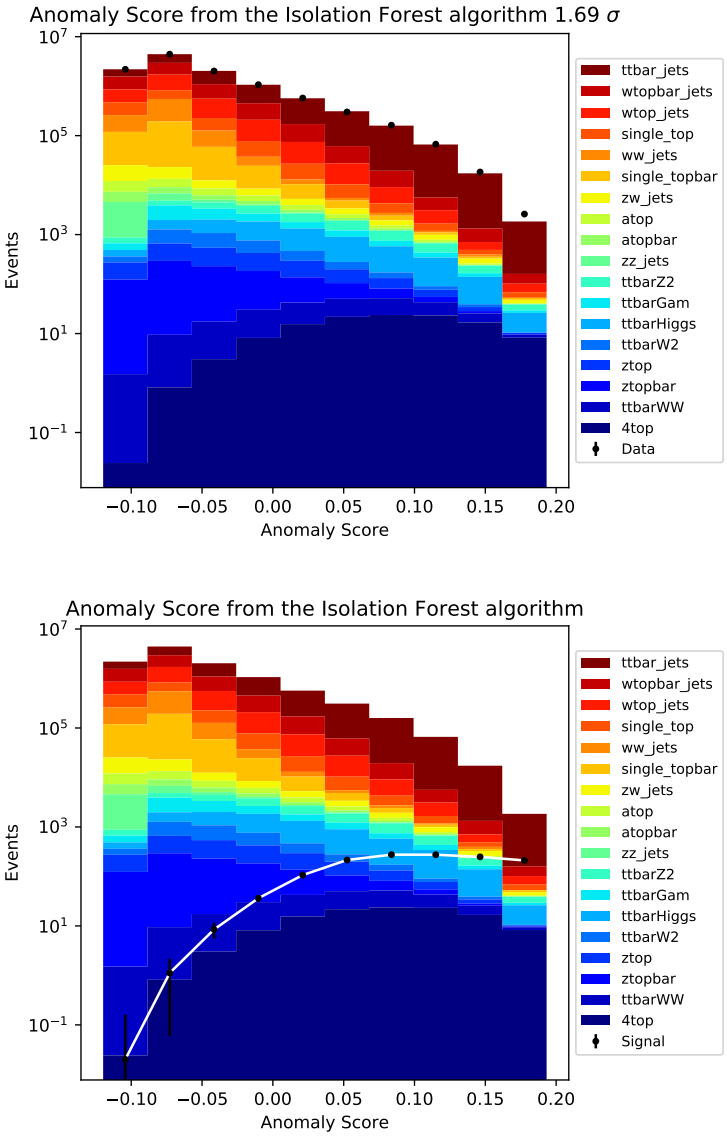

Figure 2. Results from the isolation forest tested on gluino data. On the left is a histogram of background alongside pseudo-data, and on the right is a histogram of background and pure signal. Note that the right-most histogram would not be observed in nature and is purely for demonstration.

3.2 Stop Results

The stop quark results are not as powerful as the gluino results. There is almost no overflow observed in the high anomaly region. As you can see by the right-hand side of Figure 3, the signal has very nearly the same shape as the background. This is the result of the kinematic similarity of the stop and top events in this case.

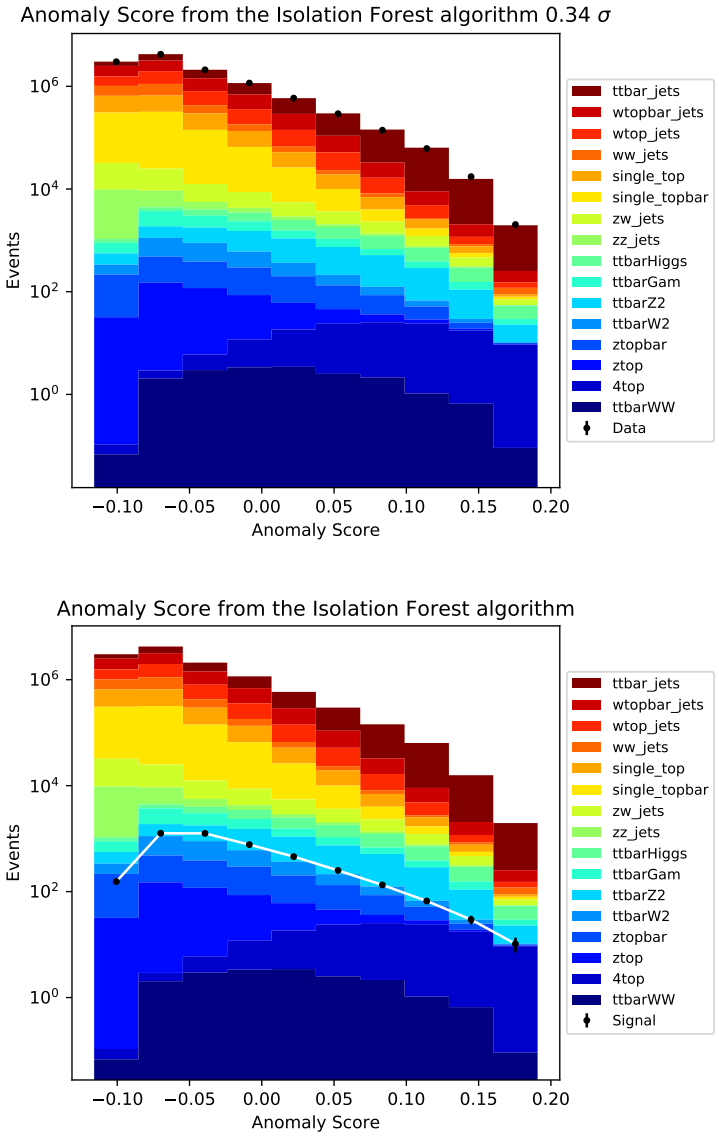

Figure 3. Results from the isolation forest tested on supersymmetric top quark data. On the left is a histogram of background alongside pseudo-data, and on the right is a histogram of background and pure signal. Note that the right-most histogram would not be observed in nature and is purely for demonstration.

4 Conclusion

Unsupervised machine learning is a powerful tool when used in performing an LHC analysis. It has been demonstrated that it is possible to find a distinction between a signal and the standard model background without making any assumptions about the signal. Unfortunately this technique alone is not enough to discover a signal on its own, however it can be used as a probing tool to analyse LHC data. Use of this technique will certainly be helpful in determining signal regions to explore further, and get one step closer to finding new physics at the LHC. Recently, alongside the Dark Machines research collective [6], we have been experimenting with further applications of this technique as one of many steps to differentiate signal from background in an unsupervised way.

References

[1] The ATLAS Collaboration et al. The ATLAS experiment at the CERN large hadron collider. *Journal of Instrumentation*, 3(08):S08003–S08003, aug 2008.

[2] ATLAS collaboration et al. Search for top-squark pair production in final states with one lepton, jets, and missing transverse momentum using 36 fb^{-1} of $\sqrt{s}$ = 13 TeV pp collision data with the ATLAS detector. *arXiv preprint arXiv:1711.11520*, 2017.

[3] Gentleman, R and Carey, VJ. Unsupervised machine learning. In *Bioconductor Case Studies*, pages 137–157. Springer, 2008.

[4] Martin, Stephen P. A supersymmetry primer. In *Perspectives on supersymmetry II*, pages 1–153. World Scientific, 2010.

[5] Liu, Fei Tony and Ting, Kai Ming and Zhou, Zhi-Hua. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[6] The Dark Machines Research Collective. https://darkmachines.org/, 2019.