Energy Flow Distributions

Unsupervised Methods in Jet Substructure Analysis

Animesh Sinha · Jai Bardhan · Kalp Shah

Submitted: July 23, 2020

Abstract In this article we present Unsupervised Machine Learning techniques on the analysis Jet Substructure using Energy Flow Polynomials as the method of choice for jet representation. We evaluate the distribution of these polynomials over different kinds of jets. These distributions are used for unsupervised probabilisitic tagging Boosted Top jets and show improved accuracy over other unsupervised techniques. Finally, we discuss anomaly detection and model-free searches for new Physics by augmenting Energy Flow Polynomials with the particle level constituents.

Keywords Jet Substructure · Energy Flow · Jet Tagging

1 Introduction

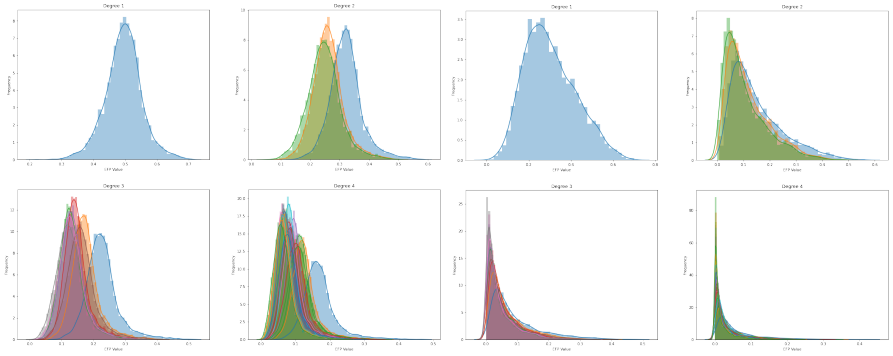
The use of Machine Learning in attempts to find new physics by probing High Energy Particle interactions has been ubiquitous, but unsupervised techiques are yet to find their place in this landscape. With vast volumes of data available from experiments like the Large Hadron Collider (LHC), it becomes increasingly desirable and ever more plausible that Unsupervised and Model-Free learning take a leading role in the hunt for new Physics.

Jet substructure is the analysis of radiation patterns and particle distributions within the collimated sprays of particles (jets) emerging from high-energy

International Institute of Information Technology
Gachibowli, Hyderabad
E-mail: animesh.sinha@research.iiit.ac.in

International Institute of Information Technology
Gachibowli, Hyderabad
E-mail: jai.bardhan@research.iiit.ac.in

International Institute of Information Technology
Gachibowli, Hyderabad
E-mail: kalp.shah@research.iiit.ac.in



(a) Histograms for QCD Jets (b) Histograms for Top Jets

Fig. 1: EFP Polynomials on Top Tagging data

collisions. [5]. It has found a central importance in many of the searches for new physics at the Large Hadron Collider (LHC) and otherwise, some of them being:

- Standard Model measurements
- Identification of boosted heavyparticles
- Discrimination of quark- from gluon- initiated jets
- Search of Beyond Standard Model particles

1.1 Energy Flow Polynomials (EFPs)

We have had several methods to represent jets as inputs to computational models, from a list of 4-vectors of constituent particles, to images of the jet itself amongst others, but either they are extremely sparse representations of the jets (as in the images) making is hard to ML models to learn something meaningful, or they have had an ordering (as in lists) not providing permutation invariance.

EFPs as a method of jet representation are both dense representations which inherently have permutation, as well as translational and rotational invariance. For any given jet, it's EFPs are Polynomials of the following form [5]:

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \dots z_{i_N} \prod_{k,l \in G} \theta_{i_k i_l} \tag{1}$$

In addition, linear models trained on these polynomials perform nearly as well as the best Neural Networks on jet tagging tasks while having over an order magnitude fewer parameters. Owing to the fact that very simple supervised models can perform well using EFPs as the input representation, it seems natural that the same would apply to unsupervised techiniques.

1.2 Latent Dirichet Allocation

Add details on what LDA does and any it's general issues, low accuracy, etc.

2 Mixture Models for Jet Tagging

2.1 Gaussian Mixture Models

As a first step we pick the boosted top tagging problem. From Figure 1, it is apparent that the jets are points N-dimentional space (here 1000 dimentional, we have limited to EFPs of degree ¡= 7) which originate from several multivariate distributions. The obvious choice to model this data is to fit a Bayesian Gaussian Mixture Model, each distribution being a multivariate Gaussian:

$$f_{pdf}(X) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{2}$$

We train the model using Expectation Maximization algorithm. The performance of the algorithm is at par with that of LDA (89.2% on the top tagging dataset).

2.2 Maxwellian Mixture Models

The performance of the GMM is marred by the fact that the distributions of EFPs are actually maxwellian. The skew of the distributions is high for the top quarks even at low degrees and increases as the degree of the EFP increases. Therefore it behooves us to model the data using the Maxwellian distribution, which follows.

We model the multivariate form of the Maxwell-Boltzmann Distribution as follows [6].

$$f_{pdf}(X) = \frac{b^{1+\frac{n}{2}} |B|^{\frac{n}{2}} \Gamma(\frac{n}{2})}{\pi^{\frac{n}{2}} \Gamma(1+\frac{n}{2})} [XBX^T] e^{-b(XBX^T)} \tag{3}$$

The parameters being learnt will be the $n \times n$ sized matrix B. X is the input vector to the model of shape $n \times 1$ and the the constant b is the determinant of the matrix B, which together with the Gamma functions serves the task of normalization.

This distribution will be trained on using the Expectation Maximization algorithm, details of the E- step and M- step will be added here, together with any improvements in performance. [9]

2.3 Performance of the Models

Following is the performance of the algorithms on the Top Tagging dataset [3].

Add accuracy figures on the QCD tagging dataset [4]

Table 1: Performance on Top Tagging [2]

Model Name	Accuracy	ROC AUC
Bayessian Maxwellian Mixture Model with EFP	UNK	UNK
Bayesian Gaussian Mixture Model with EFP	89.224%	UNK
Latent Dirichlet Allocation	89.2%	0.955
Linear Discriminant Analysis with EFP	93.2%	0.980
ParticleNet (Graph Neural Net on Point Cloud)	93.8%	0.985

2.4 Features of probabilistic tagging

While boosted top tagging is solved to a very high accuracy, the following attributes are still sought for:

- To know what confidence we have tagged a single jet with? - As seen from the histograms, some jets are very clearly in the Top or QCD domain, the overlap of the two is the set of jets where most algorithms fail. When tagging, a model should be able to output both the class label and the confidence figure.
- Resilliance to unseen input, and ability to fail gracefully if the input jet is not unlike what the model has seen before? This may be due to errors by the clustering algorithm which has clustered multiple jets together or because of an unknown decay type.

A probabilistic Bayesian generative model is the best attempt at modelling the probability with which we are tagging a jet.

Show experimental proof that we are indeed resilliant against mutliple-jet in one jet image and against unknown jet types. We need to explore what set of features help models like ParticleNet [7] perform better in the overlap zone, and it's performance relative to our confidence, as in the Orange comparative scatters.

3 Anomaly Detection using Particle Level EFPs

Write about the following here:

- How EFP preserves inner rotational symmetry therefore does not encode useless information.
- Sampling technique to escape computational complexity
- Example of how this adds to the resilliance against clustering errors.

Energy Flow Polynomials will be used as inputs to auto-encoders which are easier to construct [8].

The proof of concept for this will be shown using a similar toy model as proposed in the abstract for the LDA paper which is also doing similar Beyond-Standard-Model searches (like the toy vector-scalar boson model) [1].

4 Conclusions

Right now, Gaussian Mixture Models equalize LDA which is the state of the art of unsupervised learning on the Top Tagging dataset with 89.2%.

Pass, will come to this at the end.

References

1. Dillon, B.M., Faroughy, D.A., Kamenik, J.F.: Uncovering latent jet substructure. Physical Review D **100**(5), 056002 (2019)
2. Kasieczka, G., Plehn, T., Butter, A., Cranmer, K., Debnath, D., Dillon, B.M., Fairbairn, M., Faroughy, D.A., Fedorko, W., Gay, C., et al.: The machine learning landscape of top taggers. arXiv preprint arXiv:1902.09914 (2019)
3. Kasieczka, G., Plehn, T., Thompson, J., Russel, M.: Top quark tagging reference dataset (2019). DOI 10.5281/zenodo.2603256. URL https://doi.org/10.5281/zenodo.2603256
4. Komiske, P., Metodiev, E., Thaler, J.: Pythia8 quark and gluon jets for energy flow (2019). DOI 10.5281/zenodo.3164691. URL https://doi.org/10.5281/zenodo.3164691
5. Komiske, P.T., Metodiev, E.M., Thaler, J.: Energy flow polynomials: A complete linear basis for jet substructure. Journal of High Energy Physics **2018**(4), 13 (2018)
6. Mathai, A., Princy, T.: Multivariate and matrix-variate analogues of maxwell–boltzmann and raleigh densities. Physica A: Statistical Mechanics and its Applications **468**, 668–676 (2017)
7. Qu, H., Gouskos, L.: Particlenet: jet tagging via particle clouds. arXiv preprint arXiv:1902.08570 (2019)
8. Roy, T.S., Vijay, A.H.: A robust anomaly finder based on autoencoder. arXiv preprint arXiv:1903.02032 (2019)
9. Ueno, G., Nakamura, N., Higuchi, T., Tsuchiya, T., Machida, S., Araki, T., Saito, Y., Mukai, T.: Application of multivariate maxwellian mixture model to plasma velocity distribution function. Journal of Geophysical Research: Space Physics **106**(A11), 25655–25672 (2001)