Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Successes, Challenges and Future Outlook of Multivariate Analysis In HEP

To cite this article: Helge Voss 2015 *J. Phys.: Conf. Ser.* **608** 012058

Successes, Challenges and Future Outlook of Multivariate Analysis In HEP

Helge Voss

H. Voss, Max Plank Institute for Nuclear Physics, Heidelberg, Germany

E-mail: helge.voss@cern.ch

Abstract. Multivariate techniques using machine learning algorithms have become an integral part in many High Energy Physics data analyses. This article is intended to sketch how this development took place by pointing out a few analyses that pushed forward the exploitation of these powerful analysis techniques. This article does not focus on controversial issues like for example how systematic uncertainties can be dealt with when using such techniques, which have been widely discussed previously by other authors. The main purpose here is to point to the gain in physics reach of the physics experiments due to the adaptation of machine learning techniques and to the challenges the HEP community faces in the light a rapid development in the field of machine learning if we want to make successful use of these powerful selection and reconstruction techniques.

1. Introduction

Multivariate techniques have been used successfully since the very beginning of High Energy Physics data selection and reconstruction. The widespread use of TMVA [1], a software package which gives easy access to a variety of machine learning algorithms and which is integrated in the popular ROOT [2] analysis framework, helped to increase the knowledge about these techniques and to overcome prejudice against using such - at first glance "black box" - selection algorithms. While important issus like possible complications in addressing systematic uncertainties, apparent lack of control about the actual important figure of merit that a selection should optimise and other challenges are well known and have been reported previously by a number of authors (see for example [3]), this report focuses on the challenges of staying up to date with the rapid development in the field of machine learning in order to get the most out the very expensive accelerators and detectors.

2. Past Successes with Multivariate Analyses in HEP

This section does not intend to give a full historic overview about the use of multivariate techniques in high energy physics, but merely tries to point out a few examples which are meant to illustrate the importance of venturing into new analysis techniques in order to boost the performance of an analysis. In fact, one of the most powerful multivariate pattern recognition tools for image recognition, the human visual cortex, was used in the very early days when photos of bubble chambers were analysed. Only with the advent of computer based analysis of large data sets, many physicists reverted to simple one-dimensional, rectangular cuts to select their events of interest. These cuts are easy to formulate in a program code, their operation is easy to understand and interpret and last but not least, they are easy to communicate.

Nonetheless, there were permanent successful attempts to venture into more sophisticated selection algorithms, as for example the naïve Bayesian (Likelihood) τ-particle identification algorithm ("TAUPID") in ALEPH [4]. In the literature one can also find even earlier attempts of promoting even more complicated techniques like neural networks at dedicated HEP statistics conferences. However, it was also already then pointed out that the adaptation of these powerful classification techniques in high energy physics is 'somewhat slow' as expressed for example in the summary of [5]: *The progress of exploiting ANN in high energy physics has been somewhat slow. Partly this conservatism is due to the a misconception that ANN approaches contain an element of "black box magic" as compared to conventional approaches. I hope I have convinced the reader that this is not the case. Statistical interpretation of the answers makes the ANN approach as well-defined to use as the discriminant ones.*

As far as the author is aware, the use of multivariate pattern recognition algorithms was basically taboo in new particle searches until the LEP2 aera Higgs searches [6] when people started to apply Likelihood or Neural Network algorithms and overcame widespread scepticism by demonstrating the superior results obtained compard to classical rectangular cuts.

With the pioneering analysis of MiniBooNE that used for the first time a Boosted Decision Tree in a HEP analysis as outlined in [7], a lot of interest in has been raised in the community to exploit this "new" selection algorithm. Subsequently, boosted decision trees for sure became the favourite algorithm within the TMVA framework used in many analysis, including the first evidence for the Higgs Boson as reported by CMS [8].

Boosted Decision Trees are of course popular for a good reason. They are very robust, meaning they do not require very careful tuning in order to get close to their optimal performance, and powerful. They had for this reason sometimes been called the best 'out of the box' classifiers. Simple Decision Tree are easy to interpret and very similar to standard rectangular cuts. Each branch of the tree simply represents a cut sequence. Boosted decision trees however loose basically all the advantage of easy interpretation or understanding of the selection during the boosting sequence. They are probably as difficult to understand as Neural Networks in terms of how they place the decision boundaries in the feature space of the discriminating variables that are fed into it. Moreover, if one carefully examines the decision boundary that boosted decision trees parameterise if a certain threshold is imposed on the response variable, this is typically a very irregular, non smooth hyperplane. This is easily understandable from the fact that the Boosted Decision Trees' response is a weighted sum of rectangular volumes that are given by each of the individual trees as signal or background like areas. Even for very large number of boosting steps, this feature remains[1] as typically the boosting algorithm intrinsically doesn't give a sizeable contribution to those classifiers in the ensemble that are derived at late boosting steps. Therefore no matter how many boosting steps there are, they will not really smooth out the decision boundaries.

On the other hand, the decision boundaries from neural networks, as they are a superposition of smooth activation functions of the various nodes, are generally much smoother and for sure, after looking at a comparison of the decision boundaries from a Neural Network and a Boosted

[1] AdaBoost, the most popular boosting algorithm, for example keeps reweighting events of the original training data sample such that event, which are mis-classified by the previously trained classifier (decision tree) in the ensemble of boosted classifiers, keep getting higher weights compared to events out of the training set that we correctly classified. This results in variable distributions of the reweighted training set that keep getting more and more equal between signal and background. Decision trees trained on these reweighted data samples then obviously have a larger error rate on the training set and hence get a smaller weight in the ensemble of classification trees. Eventually the error rate approaches 50% which translates to a contribution of to the boosted ensemble classifier of zero. This behaviour is actually a very nice feature of the boosting algorithm, as it removes the necessity of actively carefully choosing the number of boosting steps as configuration parameter.

Decision Tree, most physicist would rather evaluate possible systematic on a smooth Neural Network output rather than on a very irregular one Boosted Decision Tree. Unfortunately, standard Neural Networks are much harder to train properly, particularly with limited amount of labelled training data and hence they are in reality often less performant than Boosted Decision Trees and therefore less popular.

3. New Developments in Machine Learning and Multivariate Analysis

At this point it is worth noting that at the time Boosted Decision Trees became popular in the HEP community, the hype about Boosted Decision Tree in the machine learning community had been already over and new developments in the field of neural network training have all but totally revolutionised the main areas of pattern recognition like image and speech recognition. In 2006/2007, a major breakthrough in neural network research finally allowed successful training of so called 'deep neural networks'[2]. This was achieved using a meaningful initialisation of the weights[3] connecting the neurons in each layer before starting the usual training via back-propagation. This initialisation is done by pre-training each individual layer as Restricted Boltzmann Machines [9] or as auto-encoders [10].

This next generation of neural networks proved extremely powerful and it seems that all major players in the 'industry' that apply machine learning at a large scale (e.g. Google, Facebook, Microsoft and IBM) are replacing their hand crafted and fine tuned speech and image recognition algorithms by deep neural networks which often substantially outperform their predecessors. This development can be followed in the news where major acquisitions of deep learning companies or hiring of deep learning specialist from academia by one of these companies are reported.

Also in physics, first attempts have been made to apply these new techniques and some promising results have been reported [11]. In order to understand how profound the changes are that other fields have experienced due to the advances of deep learning technologies, let's imagine we discard all of our carefully, with physics knowledge crafted high level features like invariant masses, jets, secondary vertices or impact parameters and let a large deep network learn these features by itself from simple 4-vectors as input.[4] While this might not be the best approach in HEP, the technology might well be useful for much more changes to our analyses than a simple replacement of a Boosted Decision Tree by a Deep Neural Network.

Given the sweeping success the paradigm change in terms of which type of features are used as input to the machine learning together with the deep learning network technologies, it would certainly be unwise not to study such and other possibilities of deep network applications in detail for our HEP data. Rather than lagging behind in the application of modern machine learning techniques, high energy physicist should rather be using state of the art technologies if not trying to drive the development as it has been in so many other fields.

An interesting additional aspect of these deep neural networks is that the layer-by-layer pre-training uses unlabled data, meaning it doesn't need to be known whether an event is of type "signal" or "background". This of course gives the potential to use recorded data from the experiment rather than extensive amounts of generated Monte Carlo events for this step.

A completely unsupervised training of a large deep layer neural network on HEP data might eventually perhaps not 'discover cats' as it was done by Google [12], but could be used for

[2] Neural Networks with more than 2 hidden layers are typically referred to as 'deep networks'

[3] Previously, the weights of a neural network were initialized randomly with small values, such that the network acted in the beginning much like a random linear classifier which through backpropagation was intended to quickly learn the best linear separating boundary which would then gradually be refined as the training process progresses.

[4] In speech recognition for example, over the years carefully tuned high level features (phonemes) which were previously used as input to the learning algorithms have apparently been simply replaced by a large deep neural network which learns these features.

completely model independent searches, or perhaps help in understanding the detector and hence be a vital tool for systematic studies. Obviously, the requirements in HEP are not the same as for speech or image recognition and it will probably require a substantial effort to evaluate the possibilities of these new Deep Learning strategies. Research on these Deep Neural Networks is developing at a fast pace[5] Understanding such complex algorithms and their behaviour on such data samples will certainly exceed the scope of a typical PhD thesis, let alone the ever more complex code base required for building, training and monitoring such advanced deep networks. It should be noted that particular effort is necessary to monitor and understand the working of these algorithms in order to understand all systematic aspects [6]

4. Summary

It is hard to quantify the impact of the increased use of sophisticated MVA techniques on the physics output of our current HEP experiments, but one could perhaps say that an improvement of selection performance w.r.t a simple cut based analysis of somewhere between 5% and 20% are certainly not unheard of, but rather common. This leads to a very substantial increase of the physics reach for our experiments! Compared to how much an equivalent gain in physics reach by upgrading an accelerator or detector would cost, funding of machine learning reasearch and development for physicists should be regarded as an at least equally good investment.

Simply reaching out to the other groups that do either plain machine learning research or within the context of other industrial or business applications as it is has been done so far with the funding of interdisciplinary activities which lead to initiatives like the Kaggle Higgs Challenge (https://www.kaggle.com/c/higgs-boson) is probably not enough to really make proper use of modern machine learning in HEP.

In order to deal properly with the ever increasing complexity of machine learning it will be difficult for individuals to really grasp the full potential and manage the pitfalls of those analysis techniques. While it was still possible to simply program your own likelihood selection algorithm, this was already much more challenging with standard neural networks or boosted decision trees and will be even more so with the new algorithms. But rather than simply adopting tools developed outside of (high energy) physics, the community would certainly profit considerably from an engagement into an 'in house' machine learning development and support group for HEP usage, similar as it was done with ROOT [2], which has a CERN funded core development and support team for analysis software dedicated to physicists. This would help to ensure that the data we have collected in our experiments is exploited in the best way using state of the art analysis technologies while making sure that we understand all the results to the level that we feel comfortable with as physicists.

Perhaps our data selection is less difficult than pattern recognition in other fields and we do not need such sophisticated analysis tools. However it would certainly be a loss if the possible gains would not be thoroughly studied. These new machine learning techniques are thorougly studied in the context of self driving cars for example, where 'errors' are much more severe than an under/over estimated systematic error on a physics parameter. This clearly shows that potential criticism about the use of such machine learning algorithms in High Energy Physics because we need to understand our systematic uncertainties much more thoroughly than it is the habit in other field, does not seem very compelling. Moreover it simply means that more effort is needed to be put into a deeper understanding of these state of the art techniques within

[5] Recently it has been reported for example that using so called 'rectified linear units' [13] instead of the typical sigmodial activation functions used in traditional neural networks or 'dropout' [14] again may lead to major improvement of (deep) neural network performances

[6] Note: Systemiatic uncertainties or errors never are a result of a particular training of an algorithm. The come into effect once we do not exactly understand the efficiency and background rejection of the trained algorithm, globally and for individual regions in phase space.

Journal of Physics: Conference Series **608** (2015) 012058 doi:10.1088/1742-6596/608/1/012058

the context of high energy physics.

References

[1] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," PoS A CAT 040 (2007) [physics/0703039].

[2] http://root.cern.ch

[3] L. Lyons, *"Open statistical issues in particle physics"* The Annals of Applied Statistics, Vol. 2, No. 3, (2008) 887.

[4] D. Decamp *et al.*, ALEPH Collaboration, *"Measurement of tau branching ratios"*, Z. Phys. C **54** (1992) 211.

[5] C. Peterson, *"Neural Networks and High Energy Physics"*, Proc. of International Workshop on Soft- ware Engineering, Artificial Intelligence and Expert Systems for High Energy and Nuclear Physics, Lyon Villeurbanne, France, March , 1990 , eds. D. Perret-Gallix and W. Woijcik, Editions du CRNS (Paris 1990); B. Denby, *"Neural Networks and Cellular Automata in Experimental High Energy Physics"*, Computer Physics Communications 49 (1988) 429.

[6] ALEPH, DELPHI, L3 and OPAL Collaboration, Phys. Lett. B565 (2003) 61.

[7] B.P. Roe, H-J. Yang, J. Zhu, Y. Liu, I. Stancu, G. McGregor, *"Boosted decision trees as an alternative to artificial neural networks for particle identification"*, NIMA, Vol543, (2005), 577.

[8] CMS Collaboration, *"Observation of a new boson at a mass of 125 GeV with the {CMS} experiment at the {LHC}"*, Phys Lett. B, 716 (2012) 30. bibitemAdaBoost Y. Freund and R.E. Schapire, J. of Computer and System Science 55, (1997) 119.

[9] Hinton, G. E., Osindero, S. and Teh, Y. W. *"A Fast Learning Algorithm For Deep Belief Networks"*, Neural Computation, Vol 18 (2006) 1527.

[10] Bengio, P. Lamblin, D. Popovici and H. Larochelle, *"Greedy Layer-Wise Training of Deep Networks"*, Advances in Neural Information Processing Systems 19 (NIPS06), MIT Press (2007) 153.

[11] P. Baldi, P. Sadowski, and D. Whiteson *"Searching for Exotic Particles in High-Energy Physics with Deep Learning"*, arXiv hep-ph/1402.4735v2; de Oliveira, L.P., Golling, T., *"Optimizing Deep Learning Neural Network Architectures for Jet Identification at* $\sqrt{s} = 8\,TeV$ *using the ATLAS detector"*, (In preparation).

[12] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean and A.Y. Ng, *"Building high-level features using large scale unsupervised learning"*, In International Conference on Machine Learning (ICML), Omnipress, (2012) 103; (arXiv:1112.6209v5).

[13] V. Nair and G.E. Hinton, *"Rectified linear units improve restricted boltzmann machines"*. In Proc. 27th International Conference on Machine Learning (ICML), Omnipress Madison, (2010) 807.

[14] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R.R. Salakhutdinov, *"Improving neural networks by preventing co-adaptation of feature detectors"*, CoRR (2012), arXiv:1207.0580v1.