Ph.D. thesis proposal in computer science
2013-2016

Learning to discover: supervised discrimination and unsupervised representation learning with applications in particle physics

Advisors	Balázs Kégl (CR1/CNRS,LAL/LRI), Cécile Germain (professeur UPSud, LRI)
Laboratories	Linear Accelerator Laboratory (LAL), University Paris-Sud(UPSud)/CNRS & Laboratoire de la Recherche en Intormatique (LRI), UPSud/CNRS/INRIA
Research groups	Apprentissage et Optimisation (TAO/LRI) & Machine Learning and Applied Statistics (AppStat/LAL) with support from ATLAS (LAL), ILC (LAL), and Auger (LAL)
Address	University Paris-Sud, 91898 Orsay Cedex
Mail	kegl@lri.fr, cecile.germain@lri.fr
Keywords	machine learning, classification, unsupervised representation learning, particle physics
Required background	statistics or signal processing or physics or computer science

1 Introduction

Today, machine learning methods are routinely used in high-energy particle physics to accelerate the discovery of new phenomena. Typically, standard classification algorithms are used for signal/background separation both in the online selection (trigger) step [1], and in the offline analysis [2]. The first goal of this thesis to answer some of the theoretical questions raised by these unorthodox machine learning applications, and to design new algorithms that improve the analyses. In the second theme we propose to go beyond the standard setup of "manual" feature extraction followed by classification and to investigate the applicability of recently developed techniques on unsupervised representation learning [3, 4, 5]. Both themes are motivated by concrete particle and astroparticle physics exeperiments (ATLAS@CERN) [6], the future International Linear Collider (ILC) [7], and the Pierre Auger Experiment (Auger) [8]. Data provided by these experiments will be a natural testbed for methodolgies developed in the thesis.

2 Research themes

2.1 Learning to discover

The Atlas[1] and the Compact Muon Solenoid (CMS)[2] experiments, recently claimed the discovery of a new particle that is very likely to be the Higgs boson [9, 10]. The existence of the particle was prdicted almost 50 years ago to have the role of giving mass to other elementary particles. It is also the final ingredient of the Standard Model of particle physics, ruling subatomic particles and forces. The experiment sits on the Large Hadron Collider (LHC) at CERN (the European Organization for Nuclear Research), Geneva, which began operating in 2009, after about 20 years of design and construction, and will continue operating for at least the next 10 years. The particle discovered is so far consistent with the Higgs boson, however, it has so far only been seen in three distinct decay channels. Finding it in other channels is a crucial step in proving that it is indeed the predicted Higgs boson.

The discovery of the new particle makes significant use machine learning methodologies developed in the last two decades. Typically, standard classification algorithms are used for signal/background separation in the following non-standard setup. We are given a set of simulated events $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $\mathbf{x} \in \mathbb{R}^d$ is a vector of features extracted from raw observations and $y \in \{-1, +1\}$ is $+1$ for *signal* events and $y = -1$ for *background* events. The raw features are typically obtained in detectors built at particle accelerators, and standardized/aligned features are extracted using "manually" based on background knowledge in particle physics and models of the detector.

[1] http://atlas.ch
[2] http://cms.web.cern.ch

The goal of classifier design is to find regions of the feature space where the signal is present or where it is amplified with respect to its average abundance. Once the subregion is found, we wait until the number of events in the region is significantly higher than that predicted by the pure background hypothesis. Formally (see, e.g., Eq. (97) in [11]), the goal is to find a function $f : \mathbb{R}^d \to \{-1, +1\}$ such that

$$G(f) = \sqrt{2\big((s+b)\ln(1+s/b) - s\big)} \tag{1}$$

is maximized, where

$$\begin{aligned} s &= \sum_{i=1}^{n} \mathbb{I}\{f(\mathbf{x}_i) = 1 \wedge y_i = 1\} \\ b &= \sum_{i=1}^{n} \mathbb{I}\{f(\mathbf{x}_i) = 1 \wedge y_i = -1\}. \end{aligned}$$

Whereas G is clearly related to the classical classification error

$$R(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{f(\mathbf{x}_i) \neq y_i\},$$

the two are not equivalent. A notable difference is that the *expectation* of G (and thus the optimal f) depends on the sample size n, whereas increasing sample size only decreases the *variance* of R. Nevertheless, the standard practice is to learn a discriminant function $g : \mathbb{R}^d \to \mathbb{R}$ on D using standard classification methods, and then use G only to optimize a threshold θ which defines the function f by

$$f(\mathbf{x}) = \begin{cases} +1 \text{ if } g(\mathbf{x}) > \theta \\ -1 \text{ otherwise.} \end{cases}$$

The first step of this theme is the formalization of the relationship between G and R. The goal is then to investigate whether classical classification algorithms (e.g., neural nets, boosting [12], or support vector machines [13]) can be adapted to optimize G explicitly. The final step is the experimental validation of the developed techniques. The theme is part of a project whose goal is to define and organize a series of data challenges on problems related to the selection of Higgs boson signals and to the measurement of its properties in the ATLAS experiment.

2.2 Automatic representation learning for imaging calorimeters

Future lepton colliders with center-of-mass energies of around 1 TeV will play a key role in understanding the origin of electroweak symmetry breaking. This breaking mechanism is intimately coupled to the existence of the Higgs boson or of the mass hierarchy in the fermion sector of the Standard Model of particle physics. The new generation detectors for the lepton collider will include a high-resolution pixel calorimeter to precisely measure the trajectory and the energy of particles produced by the collisions. Technically, a pixel calorimeter will produce 4D data (three spatial dimensions and deposited energy) that will allow us to determine the topology of hadronic showers to unprecedented detail.

In today's practice, calorimeter data is analyzed using highly specialized "hand-crafted" algorithms that exploit prior knowledge on physics processes. To optimize the analysis, we have recently started to use traditional machine learning algorithms to various reconstruction sub-problems (for example, data cleaning, localizing the interaction, or classifying the interaction type). To feed input to these algorithms, the raw calorimeter data first have to be digested into a vector of standardized features (for example, longitudinal a lateral profile descriptors). The classification algorithm (more specifically, MULTIBOOST) can then learn a given task based on these "manually" extracted features.

The goal of this theme is to explore the applicability of deep learning techniques for analyzing calorimeter data. Deep learning [3] is a new paradigm within the machine learning domain which has already had a major impact on natural language processing [4] and computer vision [5]. In this approach, the supervised classification step is

usually preceeded by an unsupervised pre-learning step in which a descriptive representation of the data is extracted from the raw images. The crucial advantage of the approach is to eliminate the manual feature extraction step and to automatically discover features that are relevant to different subtasks. Although the theme is directly motivated by the data analysis problem of the ILC calorimeter, the goal is to develop techniques that are also usable in the Higgs analysis (ATLAS) and in processing the water Cherenkov signal in the Auger experiment.

3 The candidate

The ideal candidate has a strong background either in computer science, statistics, or physics, and an open mind to acquire the necessary knowledge in the other two disciplines.

References

[1] V. Gligorov and M. Williams, "Efficient, reliable and fast high-level triggering using a bonsai boosted decision tree," tech. rep., arXiv:1210.6861, 2012.

[2] Aaltonen, T. et. al, "Observation of electroweak single top-quark production," *Phys. Rev. Lett.*, vol. 103, p. 092002, Aug 2009.

[3] Y. Bengio, A. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR*, vol. abs/1206.5538, 2012.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, MIT Press, 2012.

[6] ATLAS Collaboration, "ATLAS detector and physics performance technical design report," Tech. Rep. 99-15, CERN/LHCC, 1999.

[7] "International Linear Collider reference design report," 2007. `http://www.linearcollider.org/cms/?pid=10`

[8] Pierre Auger Collaboration, "Pierre Auger project design report," tech. rep., Pierre Auger Observatory, 1997.

[9] The ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys.Lett. B716*, 2012.

[10] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys.Lett. B716*, 2012.

[11] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C*, vol. 71, pp. 1–19, 2011.

[12] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119–139, 1997.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.