Thesis proposal
2010-2013

Large-scale Neyman-Pearson learning with applications in experimental physics

Advisor	Balázs Kégl
Laboratory	LRI/LAL, University Paris-Sud/CNRS
Research group	Machine Learning and Applied Statistics (AppStat) & TAO
Address	University Paris-Sud, 91898 Orsay Cedex
Mail	balazs.kegl@gmail.com
Web	http://users.web.lal.in2p3.fr/kegl

Keywords	machine learning, AdaBoost, astrophysics, signal processing
Required background	statistics or signal processing or physics or computer science

Summary

In the classical setup for pattern recognition (classification) [1], Bayes' theorem tells us that we should train on data in which the class proportions match the prior class probabilities. It turns out that there are several applications where either the prior class probabilities are unknown, or the goal is not to minimize the classification error symmetrically among classes, but to minimize the false negative (miss) rate given a preset false positive (false alarm) rate. For example, in experimental physics, virtually all trigger algorithms work in this way: we have to maximize the rate of captured events while observing a trigger rate often prescribed by bandwidth or other computational constraints. Classification algorithms have also been used in recent scientific discoveries of the electroweak production of single top quarks [2, 3]. The setup is similar: the goal is not symmetric classification but the enhancement of the signal-to-background ratio for hypothesis testing. Although this thesis will concentrate on applications in experimental physics, similar scenarios occur frequently in other application areas, such as real-time object detection [4].

Although classification algorithms are widely used in these applications, they are obviously sub-optimal since they were developed for symmetric classification. More surprisingly, there is very little theoretical work laying down the foundations for these natural applications. The only exception is a recent paper by Davenport, Baraniuk, and Scott [5] who coined the term "Neyman-Pearson learning" (NP learning) because of the obvious connection of the Neyman-Pearson lemma used in hypothesis testing. The goal of this thesis is to explore the landscape of NP learning and to develop efficient algorithms in this setup.

The plan is to strike a balance between algorithmic development and applications in experimental physics. On the algorithmic side we will concentrate on AdaBoost [6] which is one of the most influential supervised learning algorithms of the last decade. The main idea of boosting is to combine several simple models into a final predictor. The beauty of the algorithm is that the individual models do not have to be particularly good, so many suboptimal algorithms can be used as base learners. The final classifier is constructed in a stepwise fashion by adding base classifiers to a pool, one at a time, and using their weighted "vote" to determine the final classification. The principal algorithmic-theoretical goal of the thesis will be the development of a Neyman-Pearson boosting algorithm, and prove a weak-to-strong-learning boosting theorem.

There are several other aspects of NP learning that may be explored in the thesis. First, in most of the applications the problem is inherently unbalanced. In both triggers and signal-vs-background classifiers it is usual to have several orders of magnitudes more background than signal. This implies that in both cases a cascade classifier is a natural choice, and indeed, almost all the real-world triggers consist of a hand-made cascade classifier. Cascade classifiers are usually used in object detection [4], but there is no comprehensive approach to learn these classifiers automatically, so an interesting sub-goal of the thesis can be to develop such a principled approach. Second, most of the NP learning applications require to train on huge data sets, so the thesis will naturally focus on large-scale machine learning, a new paradigm proposed recently by Bottou and Bousquet [7]. Studying AdaBoost

within this framework is another possible subgoal of the thesis. Third, triggers, as object detectors, are based on multiple-instance training sets [8]: we know that an object is present in the image (or an event has occurred within a time/space window), but we do not know when and where exactly. Adapting AdaBoost to this setup is another interesting algorithmic-theoretical question.

On the applications side, we will focus on the development of the JEM-EUSO trigger. The goal of the JEM-EUSO experiment [9] is to study the properties of ultra-high energy cosmic rays by observing the particle cascade generated by the collision of the cosmic ray particle and atmospheric particles. Studying the composition, the energy, and the sources of these particles is important for understanding the universe tracing back to its origins. At the top of the spectrum, the particle energies can exceed 10^{20} eV which is roughly equivalent to the macroscopic energy of a hard-hit tennis ball! The main problem of detecting these particles is that they are very rare: there are only about 1 particle per km^2 per century over 10^{19} eV! To obtain reasonable statistics, the detector has to observe a large portion of the atmosphere. The goal of the JEM-EUSO experiment is to observe the light emitted by the air-cascade in the Earth's atmosphere from the space. JEM-EUSO will be on orbit on the Japanese Experiment Module (JEM) of the International Space Station (ISS) at the altitude of approximately 400km, starting in 2015. The sensor is a super wide-field telescope that detects high energy particles with energy above 10^{19} eV. The observational aperture of the ground area is a circle with 250 km radius which means that the instantaneous aperture of JEM-EUSO is larger than the Pierre Auger Observatory (currently the largest detector) by a factor of 50 to 250.

The focal plan of the camera consists of 200000 to 300000 pixels (Figure 1/left panel). The events correspond to linear correlations in the two spatial dimensions as well as between the spatial and time dimensions (Figure 1/right panel). The angular resolution of the camera is 0.1° whereas the time resolution is $2.5\,\mu$s. The rate of events is about 100 a day. A major challenge is the high background noise in general and also its highly variable nature (cities, lightnings, etc.).

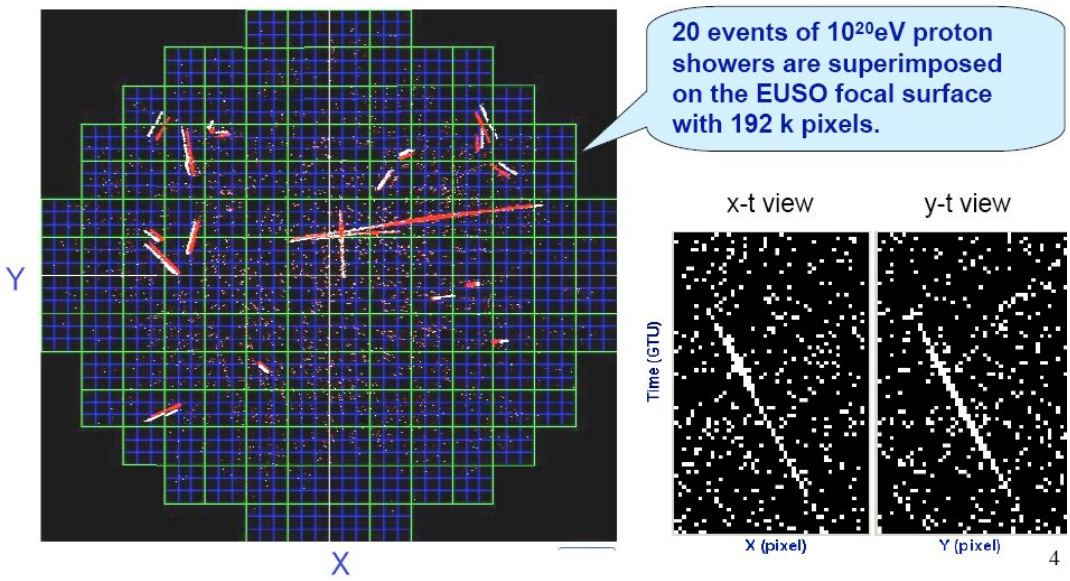

Figure 1: Images of simulated air showers over the background noise in the JEM-EUSO camera (left panel). Representing the time dimension (right panel).

The design of the on-board software faces extraordinary challenges. Its main purpose is triggering and selecting small number of candidate events (vs. background noise) which will be transmitted back to the ground. It has to be fast, computationally simple (strict limit on power consumption), and produce a low false positive rate (strict limit on transmission bandwidth) while not missing many high energy events. The collaboration is at the stage of finalizing the first and second-level triggers. The first-level trigger will look at the image using 3×3 windows, and select those that "see" a high number of photons over the background noise (where "high number" might mean only "more than a couple" per pixel; the photodetectors will have extraordinary sensitivity: they will be able to count individual

photons!). According to simulations, the trigger rate of this module will be $10^7 - 10^8\,\mathrm{s}^{-1}$. The second-level trigger will take the candidates of the first-level trigger, and filter them by looking at a small spatial and time window. It will sweep the image using a 9×9 patterns and also try to find linear correlations in time between neighboring windows. This procedure will bring down the trigger rate to about $10^3\,\mathrm{s}^{-1}$.

Our goal in this project is to design the third level trigger. Its task is to bring down the trigger rate to $0.1\,\mathrm{s}^{-1}$ which is a strict limit imposed by bandwidth limitations. For this purpose we are proposing to the collaboration a machine-learning-based approach [4] that proved itself in real-time face-detection. The approach uses AdaBoost [6] with two important additions. First, it applies a special base learner based on Haar filters, which are rectangular black-and-white wavelet-like edge detectors. The main advantage of this approach is computational efficiency which will be crucial in the proposed module: beside bandwidth limitations, the ISS has also strict limits on power consumption. The second improvement of [4] is the design of a cascade-classifier, similar in spirit to the multi-level trigger used in most of the rare-event detectors. Low-level small-complexity classifiers get rid of most of the background images very fast, and pipeline only a few candidates to higher level classifiers. While this is more complex method at the training phase, it accelerates the detection procedure by orders of magnitudes. This, again, will be a very important technical advantage in our application.

The candidate is expected to adapt the multiboost software package to the application and tune the software using simulated data. The main technical difficulty in adapting [4]'s methodology to the third-level trigger of JEM-EUSO is the additional time dimension and the fact that we have strict limits on the false positive rate.

References

[1] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.

[2] T. Aaltonen et al., "Observation of electroweak single top-quark production," *Physical Review Letters*, vol. 103, no. 9, 2009.

[3] V.M. Abazov et al., "Observation of single top-quark production," *Physical Review Letters*, vol. 103, no. 9, 2009.

[4] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.

[5] M. Davenport, R. Baraniuk, and C. Scott, "Tuning support vector machines for minimax and neyman-pearson classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, 2010.

[6] 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.

[7] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems*, vol. 20, pp. 161–168, 2008.

[8] P. Viola, J. Platt, and Z. C., "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems*, vol. 18, pp. 1417–1424., 2006.

[9] Y. Takizawa et al., "JEM-EUSO: Extreme Universe Space Observatory on JEM/ISS," *Nuclear Physics B - Proceedings Supplements*, vol. 166, pp. 72–76, 2007.