Discovering Neutrinos Through Data Analytics

Mathis Börner(✉), Wolfgang Rhode, Tim Ruhe,
for the IceCube Collaboration, and Katharina Morik

TU Dortmund University, Experimental Physics, Computer Science,
Otto-Hahn-Str. 12, 44227 Dortmund, Germany
{mathis.boerner,wolfgang.rhode,tim.ruhe,katharina.morik}@tu-dortmund.de
http://sfb876.tu-dortmund.de/SPP/sfb876-c3.html,
http://icecube.wisc.edu

Abstract. Astrophysical experiments produce Big Data which need efficient and effective data analytics. In this paper we present a general data analysis process which has been successfully applied to data from IceCube, a cubic kilometer neutrino detector located at the geographic South Pole.

The goal of the analysis is to separate neutrinos from atmospheric muons within the data to determine the muon neutrino energy spectrum. The presented process covers straight cuts, variable selection, classification, and unfolding. A major challenge in the separation is the unbalanced dataset. The expected signal to background ratio in the initial data (trigger level) is roughly $1{:}10^6$. The overall process was embedded in a multi-fold cross-validation to control its performance. A subsequent regularized unfolding yields the sought after neutrino energy spectrum.

Keywords: Neutrinos · IceCube · Machine learning · Random forest · Feature selection · Cross-validation · Signal and background separation

1 Introduction

IceCube is a neutrino detector located at the geographic South Pole with an instrumented volume of a cubic kilometer [1]. The detector consists of 86 strings at depths between 1450 m and 2450 m. Each string holds 60 digital optical modules (DOMs). The DOMs are designed to measure light and send a digitized signal to the surface. The purpose of the instrumentation is to measure Cherenkov light emitted by charged particles propagating through natural ice. The appearance of such a particle in the detector is referred to as an event. There are two types of events: events induced by neutrinos interacting in (or close to) the detector and events from muons which are produced in cosmic ray air showers in the atmosphere. Since only the neutrino spectrum is sought after the separation between atmospheric muon events and neutrino events is essential. Here, we summarize the a separation process implemented in RapidMiner [5] (based on [2]) and the subsequent unfolding.

This paper is based on work with the IceCube collaboration [3] and work in project C3 of the Collaborative Research Center SFB 876 which is funded by the DFG.

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 208–212, 2015.
DOI: 10.1007/978-3-319-23461-8_15

2 Selection of Neutrinos Events

Data for this analysis were taken, when the detector was under construction and consisted of 59 strings (IC59). The analysis faces two major challenges: In the initial data for each neutrino event 10^6 atmospheric muon events are detected and the rate of neutrinos decreases with the energy, proportional to $\sim E^{-3.7}$. The initial data rate is lowered, thereby complex reconstructions become feasible. This preselection also improves the signal to background ratio to roughly 1:1000.

The signature of atmospheric muons entering the detector shows no topological difference from an event induced by a muon neutrino. The approach is based on the fact that muons can only penetrate a few kilometers of ice while neutrinos can travel even through the earth's core. Hence, the presented approach only looks for events going upwards in the detector, towards the surface. Based on misreconstruction atmospheric muons dominate the data instead of muon neutrinos even for events with a reconstructed up-going track. To select only muon neutrino events a separation between well- and misreconstructed events needs to be conducted.

2.1 Data Preprocessing

The preprocessing consisted of two cuts. The first cut was applied on the reconstructed zenith angle to select up-going events. A second cut was applied on the *line fit* velocity [1] to reject spherical events, a topology that does not occur in high quality muon neutrino events. Both cuts were optimized simultaneously with respect to background rejection and signal efficiency. These two cuts rejected 91.4% of the background and retained 57.1% of the signal.

2.2 Variable Selection

Because not all variables are equally well suited for the event selection a representation in fewer dimensions needs to be found. Therefore, the Minimum Redundancy Maximum Relevance (MRMR) algorithm in the fast implementation of [6] was used for the selection of variables. Twenty-five variables were selected as this number shows a reasonable reduction of dimensions without losing too much information while showing stable behavior in the selection (detailed list of variables in [3]).

2.3 Performance of the Random Forest

From the machine learning point of view the event selection can be formalized in terms of a classification task with the classes *signal* (atmospheric neutrinos) and *background* (atmospheric muons). A Random Forest [8] was chosen as the machine learning algorithm. Training and testing were carried out in a standard five-fold cross-validation.

[1] Speed of the reconstructed event in the detector.

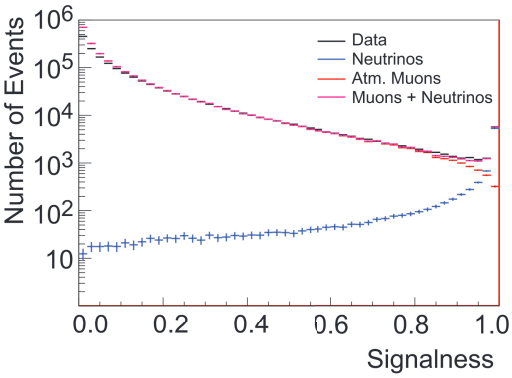

Fig. 1. Signalness (ratio of trees classifying an event as signal) distributions for data events (black) in comparison to the distributions of simulated events. Simulated signal events (Neutrinos) are depicted in blue, background events (atmospheric Muons) in magenta. The sum of the simulated events is shown in red.

The results of the Random Forest for simulated events and experimental data are shown in Fig. 1. For the analysis a strict cut of $S = 1$ was chosen. The good match between experimental data and simulated events indicates a stable performance of the forest. The errors for the distributions of the simulated events were obtained via cross-validation. The size of the errors is reasonably low and indicates a stable classification without any problems due to statistical fluctuations in the training events.

The purity of the final neutrino event sample was estimated to be $99.59^{+0.36}_{-0.37}\%$, while 18.2% of the signal is retained and 99.9999% of the background rejection is rejected.

3 Unfolding and the Resulting Energy Spectrum

The measurement of the neutrino energy is a so-called inverse problem. For this analysis the neutrino energy spectrum has to be reconstructed from measurements of the muons they induced. It can be expressed in an integral equation

$$g(y) = \int A(E, y) f(E) \, \mathrm{d}E, \tag{1}$$

where $f(E)$ is the sought-after energy spectrum, $g(y)$ the distribution of measured variables and $A(E, y)$ a function describing the whole process from the production of the neutrino until the measurement in the detector.

To solve the integral equation, a regularized unfolding method was chosen (TRUEE [7]). The approach allows us to use up to three different variables. In this analysis the three variables were: the total amount of charge in the DOMs, number of unscattered photons and the length of the track from unscattered photons.

The resulting spectrum, related measurements, and theoretical predictions are shown in Fig. 2. It shows agreement both with related measurements and theoretical predictions.

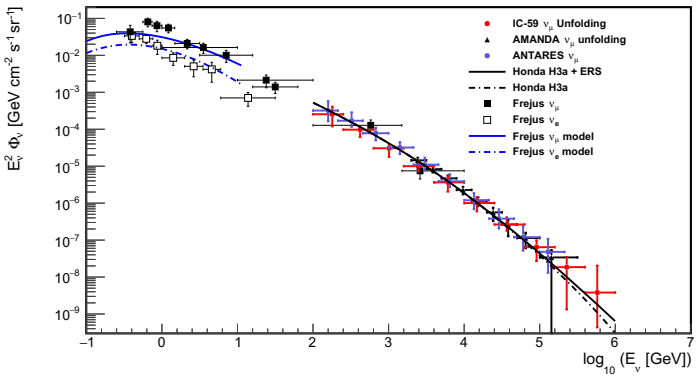

Fig. 2. The results of the analysis presented here are shown as red circles. Other measurements are depicted in black squares, hollow squares, black triangles, green triangles and blue. The curves shown originate from theoretical predictions. [3]

4 Summary and Results

This paper presents a data mining process that was successfully applied to and validated on data of the IceCube detector in its 59-string configuration. It was able to obtain 27771 atmospheric neutrino candidates in 346 days of IC59. The event selection method increased the neutrino rate from 49.3 neutrino events per day [4] to 80.3 neutrino events per day. The purity of the final neutrino sample was estimated to be $99.59^{+0.36}_{-0.37}\%$. The subsequent unfolding shows good agreement with prior measurements and extends the spectrum to energies never measured before.

References

1. Achterberg, A., et al.: First Year Performance of the IceCube Neutrino Telescope. Astroparticle Physics **26**, 155 (2006)
2. Ruhe, T., Morik, K., Rhode, W.: Application of RapidMiner in Neutrino Astronomy. In: Hofmann, M., Klinkenber, R. (eds.) RapidMiner: Data Mining Use Cases and Business Analytics Applications. CRC Press Book (2013)
3. Aartsen, M.G., et al.: Development of a general analysis and unfolding scheme and its application to measure the energy spectrum of atmospheric neutrinos with IceCube. The European Physical Journal C **3**, 75–116 (2015)
4. Abbasi, R., et al.: Measurement of the atmospheric neutrino energy spectrum from 100 GeV to 400 TeV with IceCube. Phys. Rev. D **83**(1), 012001(19) (2011)
5. Mierswa, I., Klinkenberg, R., Fischer, S., Ritthoff, O.: A Flexible Platform for Knowledge Discovery Experiments: YALE - Yet Another Learning Environment (2000)
6. Schowe, B., Morik, K.: Fast-ensembles of minimum redundancy feature selection. In: Okun, O., Valentini, G., Re, M. (eds.) Ensembles in Machine Learning Applications. SCI, vol. 373, pp. 75–95. Springer, Heidelberg (2011)

7. Mileke, N., et al.: Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics. Nuclear Instruments and Methods in Physics Research A **697**, 133–147 (2013)
8. Breiman, L.: Random Forests. Mach. Learn. **45**(1), 5–32 (2001)