Astronomical Data Analysis Software and Systems XXI
ASP Conference Series, Vol. 461
Pascal Ballester, Daniel Egret, and Nuria P. F. Lorente, eds.

Data Mining IceCube

Tim Ruhe[1] for the IceCube collaboration, and Katharina Morik[2]

[1] *Department of Physics 5b TU Dortmund University, 44221 Dortmund, Germany*

[2] *Department of Computer Science TU Dortmund University, 44221 Dortmund, Germany*

Abstract. IceCube is a 1 km^3 scale neutrino telescope located at the geographic South Pole. The large number of reconstructed attributes as well as the small signal to background ratio in an atmospheric neutrino analysis makes IceCube well suited for a detailed study within the scope of machine learning. A systematic study intended to improve the event selection was carried out including a detailed feature selection using MRMR as well as the training and testing of a Random Forest. Finally, the forest was applied on IceCube data. A good agreement between data and Monte Carlo expectations has been observed.

1. Introduction

The IceCube neutrino telescope (Ahrens et al. 2004) was completed in December 2010 at the geographic South Pole. There are 5160 Digital Optical Modules (DOMs) mounted on 86 vertical cables (strings) forming a three dimensional array of photosensors. The spatial distance between individual strings is 125 m. IceCube DOMs are buried at depths between 1450 m and 2450 m corresponding to an instrumented volume of 1 km^3. The spacing of individual DOMs on a string is 17 m (Ahrens et al. 2004; DeYoung 2009).

The detection principle of IceCube is based on Cherenkov light emitted by charged leptons created via the interaction of neutrinos with nuclei in the ice or the bedrock. Atmospheric neutrinos are produced in extended air showers where cosmic rays interact with nuclei of the Earth's atmosphere. Within these interactions mainly pions and kaons are produced which then subsequently decay into muons and neutrinos (Honda et al. 1995). The measurement of the atmospheric neutrino spectrum is hindered by a dominant background of atmospheric muons. A rejection of atmospheric muons can be achieved by selecting upward going tracks only since the Earth is opaque to muons. However, a small fraction of atmospheric muons is still mis-reconstructed as upward going.

For the starting point of this analysis (the so-called Level 3) where many advanced reconstruction algorithms have already been run and the dominant part of the atmospheric muons has already been removed, we expect $N_{back} \approx 9.699 \times 10^6$ background events and $N_{sig} \approx 1.5788 \times 10^4$ signal events in 33.28 days of IceCube in the 59-string configuration. This corresponds to a signal to background ratio of $R = 1.63 \times 10^{-3}$.

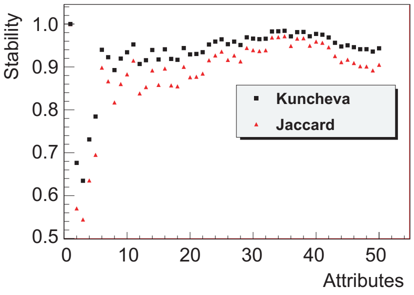

Figure 1. Stability estimation for the MRMR Feature Selection depicting the Jaccard and Kuncheva's index. The stability of the feature selection goes into saturation as the number of attributes increases. For a number of attributes ≥ 20 both stability measures lie above 0.85.

Approximately 2600 reconstructed attributes were available at Level 3. The low signal to background ratio in combination with the large number of attributes makes this task well suited for a detailed study within the scope of machine learning.

2. Feature Selection and Feature Selection Stability

This work is based on a sample where cuts of $v_{LineFit} > 0.19$ and $\theta_{Zenith} > 88^{\circ}$ were already applied previously in order to further reject the muonic background. Furthermore, we reduced the number of attributes entering our final feature selection by excluding those that were known to not contribute to the improvement of the selection. This pre-selection of attributes reduced the number of attributes entering the final selection to 477.

A Maximum Relevance Minimum Redundancy (MRMR) (Ding & Peng 2003; Schowe & Morik 2010) algorithm embedded within the F S E (Schowe 2010) for R M (Mierswa et al. 2006) was used for feature selection. Simulated events from C (Heck et al. 1998) were used as background. Simulated events from the IceCube neutrino generator N were used as signal. The machine learning environment R M (Mierswa et al. 2006) was used throughout the study.

To obtain a reliable set of features an estimation of the stability of the feature selection is crucial. The F S S V , also included in the F S E for R M , was used to estimate the stability and the outcome is depicted in Figure 1. The Jaccard index is depicted by triangles, whereas squares represent Kuncheva's index (Kuncheva 2007). Figure 1 clearly shows that MRMR can be considered stable on IceCube Monte Carlo simulations if the considered number of attributes in the selection is $n_{Attributes} \geq 20$.

3. Training and Application of a Random Forest

Figure 2 (a) shows the output of the Random Forest after a 5-fold cross validation. Within this cross validation 3.8×10^{5} simulated background and 7×10^{4} simulated signal events were used. In order to avoid overtraining the number of events used for training was limited to 2.8×10^{4} for each class. The number of trees in the forest was

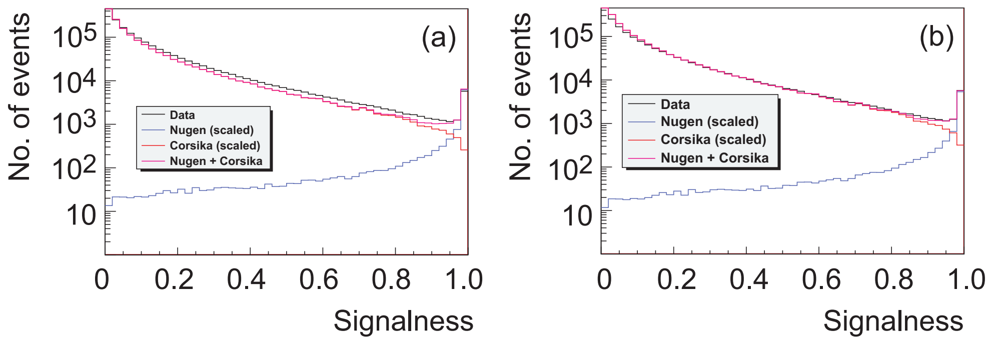

Figure 2. Random Forest score (signalness) for simulated signal and background events as well as for data. When the total number of MC events is scaled to match data in the absolute number of events a data/MC mismatch is observed for signalness > 0.2 (a). When the total number of MC however, is scaled up by 23% to match the data for signalness > 0.2 a mismatch is only observed for small signalness values (b).

chosen to be $n_{trees} = 500$. The simulation was scaled to the expected number of events for each class in real data.

In Figure 2 (a) a data/MC mismatch for a signalness $s \geq 0.2$ is observed, which would in turn lead to an underestimation of the remaining muonic background. To achieve a realistic background estimate the Corsika events are rescaled by a factor of 1.23 such that they match the distribution of data for $s \geq 0.2$. This leads to a data/MC mismatch only in the low signalness region (Figure 2 (b)).

Due to the small error bars on the expected number of signal and background events for individual cuts the performance of the forest can be considered stable (see Table 1). No indications of overtraining were observed within the cross validation. Note however that the large error bars on the number of expected background events are due to small statistics when cuts in the high signalness regions are applied.

The performance of the forest on data lies within the range expected from the cross validation. Only for a signalness cut of $s = 1.0$ one finds an under-fluctuation of 96 % of the expected number of events.

The last two columns of Table 1 show the expected purity of the final neutrino sample under the assumption that the number of background events is as expected from the mean of the cross validation (column 6) and as a worst case scenario (column 7). One finds that in both cases a purity well above 95 % can be achieved for the cuts listed in Table 1. Note that the error bars shown in Table 1 represent the statistical errors derived from the cross validation only. Systematic errors, like ice properties, DOM sensitivity or uncertainties regarding the production of neutrinos in the atmosphere (spectral index, flux normalization, π/K-ratio) have not yet been included.

4. Summary and Outlook

An investigation within the scope of machine learning was carried out using IceCube data and Monte Carlo simulations for the detector in the 59-string configuration. The MRMR algorithm implemented in the F S E (Schowe 2010) for the data mining environment R M (Mierswa et al. 2006) was used for feature

Cut	Est. Back.	Est. Sig.	Sum	Data	Pur.[%]	Pur.(worst case) [%]
0.990	114 ± 57	4817 ± 44	4931 ± 64	4988	97.7	96.7
0.992	98 ± 37	4633 ± 43	4731 ± 57	4757	97.9	97.1
0.994	71 ± 37	4414 ± 41	4485 ± 55	4476	98.4	97.6
0.996	60 ± 32	4122 ± 32	4182 ± 45	4134	98.5	97.8
0.998	22 ± 20	3695 ± 44	3717 ± 50	3638	99.4	98.8
1.000	5 ± 11	2932 ± 33	2937 ± 35	2833	99.8	99.4

Table 1. Estimated number of signal and background events as well as the estimated purity after an application of cuts on the signalness. The number of data events yielded for individual cuts is shown as well. The error bars represent the statistical error derived from the cross validation only. Systematic errors have not yet been included.

selection. The selection was found to be stable if the number of attributes considered exceeded 20. A Random Forest was trained and tested on Monte Carlo simulations using a 5-fold cross validation. Finally, the forest was applied on real data and the outcome of this application was compared to expectations derived from Monte Carlo simulations. A good agreement between data and Monte Carlo expectations was observed. We would like to note however, that the results presented in this paper are preliminary since the Random Forest has not yet been optimized. By doing so in the near future we hope to obtain even better results.

Acknowledgments. Part of the work on this paper has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 ”Providing Information by Resource Constrained Analysis”, project C3. We also acknowledge the support from the German Ministry of Education and Research (BMBF).

References

Ahrens, J., et al. 2004, Astroparticle Physics, 20, 507. arXiv:astro-ph/0305196

DeYoung, T. 2009, Modern Physics Letters A, 24, 1543. 0906.4530

Ding, C. H. Q., & Peng, H. 2003, in 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003), 11-14 August 2003, Stanford, CA, USA (IEEE Computer Society), 523

Heck, D., Knapp, J., Capdevielle, J. N., Schatz, G., & Thouw, T. 1998, CORSIKA: a Monte Carlo code to simulate extensive air showers. (Forschungszentrum Karlsruhe GmbH)

Honda, M., Kajita, T., Kasahara, K., & Midorikawa, S. 1995, Phys.Rev.D, 52, 4985. arXiv:hep-ph/9503439

Kuncheva, L. I. 2007, in Proceedings of the 25th IASTED international Multi-Conference (Innsbruck,Austria)

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. 2006, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), edited by T. Eliassi-Rad, L. H. Ungar, M. Craven, & D. Gunopulos (New York, USA: ACM Press), 935

Schowe, B. 2010, http://sourceforge.net/projects/rm-featslext

Schowe, B., & Morik, K. 2010, in Workshop on Supervised and Unsupervised Ensemble Methods and their Applications - SUEMA 2010, edited by O. Okun, G. Valentini, & M. Re. URL http://suema10.dsi.unimi.it/suemafiles/SUEMA10_proceedings.pdf