Data Mining on Ice

Tim Ruhe for the IceCube Collaboration[†], Katharina Morik and Benjamin Schowe

Abstract In an atmospheric neutrino analysis for IceCube's 59-string configuration, the impact of detailed feature selection on the performance of machine learning algorithms has been investigated. Feature selection is guided by the principle of maximum relevance and minimum redundancy. A Random Forest was studied as an example of a more complex learner. Benchmarks were obtained using the simpler learners k-NN and Naive Bayes. Furthermore, a Random Forest was trained and tested in a 5-fold cross validation using 3.5×10^4 simulated signal and 3.5×10^4 simulated background events.

1 Introduction

The IceCube neutrino telescope [4] 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 strings 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 [4, 6, 12].
A low energy extension called DeepCore [6, 12] is installed in the center of the de-

[†] For a complete author list see: http://www.icecube.wisc.edu

Tim Ruhe for the IceCube collaboration
Department of Physics TU Dortmund University, e-mail: tim.ruhe@udo.edu

Katharina Morik
Department of Computer Science TU Dortmund University, e-mail: katharina.morik@tu-dortmund.de

Benjamin Schowe
Department of Computer Science TU Dortmund University, e-mail: benjamin.schowe@tu-dortmund.de

tector. The IceTop [15] air shower array is located on top of the in-ice part of the detector.
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 [9]. Atmospheric neutrinos can be distinguished from an astrophysical flux by their much softer energy spectrum which follows a power law $\frac{dN}{dE}_{atmo} \propto E^{-3.7}$ [9].
The measurement of the atmospheric neutrino spectrum, however, is hindered by a dominant background of atmospheric muons also produced in cosmic ray air showers. Although the detector is shielded by the antarctic ice cap, atmospheric muons enter the detector due to their high energies. 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 misreconstructed 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 $N_{sig} \approx 1.418 \times 10^4$ signal events in 33.28 days of IceCube in the 59-string configuration. This corresponds to a signal to background ratio of $R = 1.46 \times 10^{-3}$. Approximately 2600 reconstructed attributes where available at Level 3.
The remaining background of atmospheric muons can further be reduced by applying straight cuts [1] or by the use of machine learning algorithms [2]. The low signal to background ratio in combination with the large number of attributes available at Level 3 makes this task well suited for a detailed study within the scope of machine learning. The selection of a subset of attributes is as important as the test of different classification algorithms if we want to obtain good results.
Since Boosted Decision Trees have already been used successfully in atmospheric neutrino analyses [2], we tested a Random Forest [5] as an example for a more sophisticated algorithm. Benchmarks were obtained using k-NN and Naive Bayes.

2 Feature Selection and event classification

Prior to our studies precuts where applied on $v_{LineFit} > 0.19$ and $\theta_{Zenith} > 88^\circ$ in order to further reject the muonic background. Furthermore, we reduced the number of attributes entering our final attribute selection by hand exluding attributes that were known to be useless, redundant or a source of a potential bias. This preselection of attributes reduced the number of attributes entering the final selection to 477. This reduced the required memory and computing time dramatically.
A Maximum Relevance Minimum Redundancy (MRMR) [7, 13] algorithm embedded within the FEATURE SELECTION EXTENSION [14] for RAPIDMINER [11] was used for feature selection. Simulated events from CORSIKA [8] were used as back-

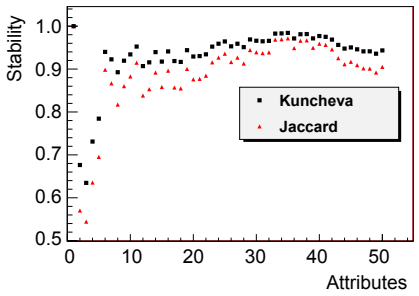

Fig. 1 Stability estimation for the MRMR Feature Selection depicting the Jaccard and Kuncheva's index. The stability of the feature selection goes into saturation as the number of attributes increases. For a number of attributes ≥ 20 both stability measures lie well above 0.9. One should note that both indices reach the maximum of 1.0 if only one attribute is selected indicating that there is one single best attribute for the separation of signal and background.

ground. Simulated events from the IceCube neutrino generator NUGEN were used as signal. The machine learning environment RAPIDMINER [11] was used throughout the study.

2.1 Feature Selection Stability

It is quite important that the feature selection given one part of the data does not differ too much from the selection given another part of the data. The ideal is, that for all parts of the data, the same features were selected. In this case the feature selection operator is called "stable". Stability is measured in terms of the Jaccard index, for instance.

Figure 1 depicts the stability of MRMR. The FEATURE SELECTION STABILITY VALIDATION, also included in the FEATURE SELECTION EXTENSION for RAPIDMINER, was used to estimate the stability. Within the FEATURE SELECTION STABILITY VALIDATION, MRMR was run in a 10-fold cross validation which itself was located in a loop that increased the number of attributes to be considered by MRMR by one per iteration.

The Jaccard index is depicted by triangles, whereas squares represent Kuncheva's index [10]. The calculation of the stability was carried out by computing the pairwise average of all subsets drawn in the cross validation.

Figure 1 shows that the stability of the feature selection rises rapidly as the number of attributes increases. For a number of attributes $n_{Attributes} \geq 10$, the stability of the MRMR selection becomes saturated and is well above 0.9 for a number of attributes exceeding 20. One should note that both indices reach their maximum value of 1.0 if only one attribute is considered. That means there is one single best attribute for the separation of signal and background in IceCube. This does not say that this sin-

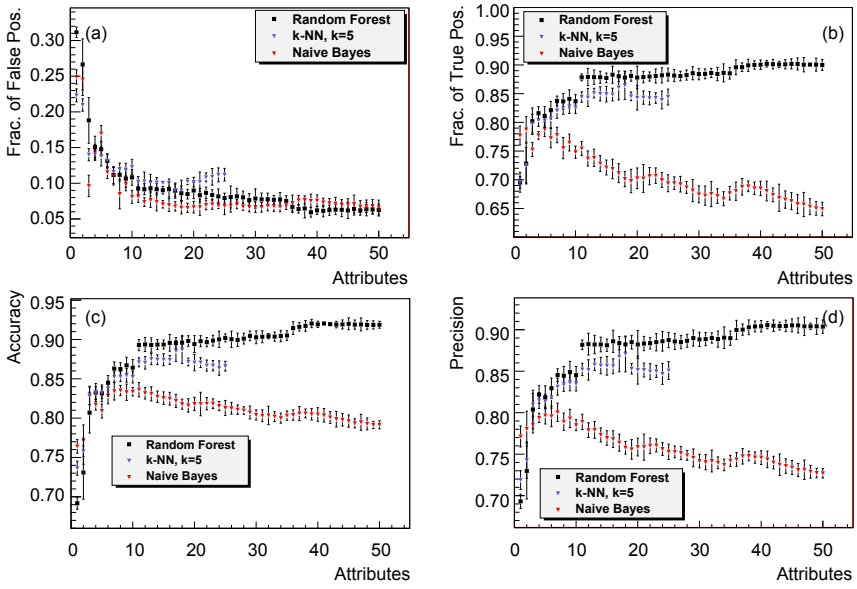

Fig. 2 Performance of the MRMR selection for 3 different learners (Random Forest, Naive Bayes, k-NN). In (a) and (b) the fractions of true and false positives are shown as a function of the number of attributes. Random Forest and Naive Bayes have a comparable performance with respect to false positives. With respect to true positives, however, the Random Forest outperformes Naive Bayes as well as k-NN with $k = 5$. In (c) and (d) accuracy and precision are shown as a function of the number of attributes. Random Forest performs better than Naive Bayes and k-NN with $k = 5$.

gle best attribute is sufficient for the separation task, but that it is a feature found relevant in the majority parts of the data. 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 $n_{Attributes} \geq 20$.

2.2 Performance

Figure 2 shows the performance of Naive Bayes, k-NN and a Random Forest after an MRMR selection as a function of the number of attributes. All learners were trained and evaluated in a 10-fold cross validation using 10^4 signal and 10^4 background events respectively. For k-NN a weighted vote and a mixed Euclidean distance was used. The number of neighbors for k-NN was chosen to $k = 5$. The Random Forest was trained using the Random Forest from the RAPIDMINER Weka package. The number of trees n_{trees} was matched to the number of attributes $n_{Attributes}$ in every iteration such that $n_{trees} = 10 \times n_{Attributes}$.

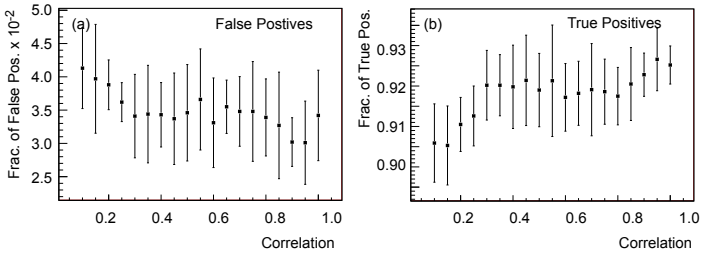

Fig. 3 Attributes, which are correlated with at least one other attribute were excluded prior to the MRMR selection, if their correlation coefficient ρ exceeded a certain value (x-axis). The dependence of the performance on this removement was studied. The performance is best for $\rho = 0.95$. The differences compared to $\rho = 0.9$, however, are negligible but show smaller errorbars which indicates a higher stability of the forest.

Figure 2 (a) shows the fraction of false positives as a function of the number of attributes. One finds that the fraction of false positives rises rapidly if the number of attributes becomes ≤ 10. While for the Naive Bayes classifier a minimum is reached around $n_{Attributes} \approx 18$ the values continue to decrease for the Random Forest. For k-NN a minimum around $n_{Attributes} \approx 18$ is reached as well. The shape of the ongoing curve, however, behaves differently from that observed for the Naive Bayes case. It rises much steeper. For k-NN the process stopped at $n_{Attributes} = 25$ as the required memory exceeded the available resources.
Figure 2 (b) depicts the number of true positives as a function of the number of attributes. For k-NN and Random Forest the curve rises rapidly and reaches a saturation around $n_{Attributes} \geq 10$. For Naive Bayes a peak is found at $n_{Attributes} \approx 5$. For $n_{Attributes} > 5$ the number of true positives decreases. A similar behavior was found for accuracy and precision shown in figure 2 (c) and (d) respectively.
A comparison of the performance of all three learners shows that the use of a Random Forest in an IceCube analysis is justified by the better performance compared to that of more simple classifiers.

2.3 Removing further correlations

A visual inspection of the attributes selected by MRMR revealed that some of the features selected were still highly correlated. The dependence of the performance of Random Forest on this correlation was investigated. As a consequence a correlation filter was applied. Within this filter one of two attributes is removed prior to MRMR if their correlation coefficient exceeds a user specified value. The correlation coefficient was varied in order to investigate the dependence of the performance of Random Forest on this coefficient.
All forests were trained and evaluated using a 10-fold cross validation with 10^4 sim-

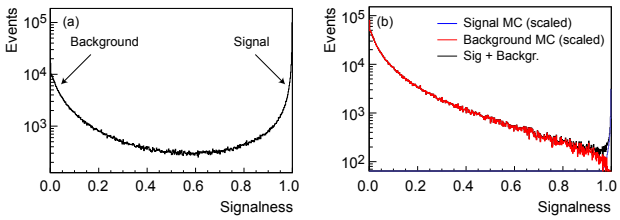

Fig. 4 a): Output of the Random Forest after a 5-fold cross validation. Two peaks are observed at $s = 0.0$ and $s = 1.0$ with the first one being mainly background and the second one mainly signal. b): Output of the Random Forest scaled to the number of signal and background events expected in real data. Again two peaks are observed with the signallike peak being significantly smaller due to the small signal to background ratio.

ulated signal events and 10^4 simulated background events. The results are depicted in figure 3. Figure 3 (a) shows that the fraction of false positives decreases as the correlation coefficient ρ of the removed attributes increases. A minimum is reached at $\rho = 0.95$. The fraction of false positives for $\rho = 0.9$, however, shows only a negligible deviation from the minimum but a much smaller error bar indicating a more stable performance of the forest.
From figure 3 (b) one finds that the fraction of true positives increases as the correlation of the removed attributes incereases reaching a maximum at $\rho = 0.95$. The fraction of true positives for $\rho = 0.9$ shows only small deviations from the optimum value but again a smaller error bar. This indicates a more stable performance of the forest.
Taking into account the negligible deviations from the optimum for both measures in figure 3 and the more stable performance one finds that attributes with $\rho \geq 0.9$ should be removed prior to an MRMR selection.

2.4 Training and Testing a Random Forest

As a result of our previous investigations, a Random Forest was trained and tested using the attributes derived in MRMR feature selection. The training and testing was carried out in a 5-fold cross validation using 3.4×10^5 simulated background events and 3.4×10^5 simulated signal events. The number of trees in the forest was chosen to $n_{trees} = 500$. To prevent overfitting the number of events used for training was chosen to 28000 signal and background events respectively.
The outcome of the testing is presented in figure 4 where figure 4 (a) depicts the signalness assigned to individual events by the forest. Figure 4 (a) shows two peaks of the signalness s. The first peak around $s = 0.0$ can be associated with background events whereas the peak at $s = 1.0$ can be associated with signal events.
Figure 4 (b) on the other hand shows the signalness of the individual events scaled to the expected number of signal and background events in real data. Again, two

peaks are found at $s = 0.0$ and at $s = 1.0$ where the peak at $s = 1.0$ is significantly smaller due to the low signal to background ratio.
By applying an additional cut on the signalness the number of background events in the final sample can be reduced while the purity of the neutrino sample increases. A couple of cuts on the signalness where applied and the remaining background as well as the purity of the final sample was computed. The outcome of this calculation is presented in table 1. The number of background events was computed in a rather conservative estimate using the upper limit of the errorbar calculated in the cross validation. Table 1 tells us that a purity well above $P = 95\%$ can routinely be achieved. Note, that for $s \geq 0.998$ and $s = 1.000$ the expected number of signal events is > 3000 and > 3800.
In addition, so far no optimization procedure was carried out on the Random Forest. By doing so in the near future we hope to achieve even better.
However, we would like to note that these numbers were calculated on the basis of Monte Carlo simulations only and might be subject to changes when applying the procedure on real data. Changes in event numbers for signal and background might occur due to data MC mismatches or due to uncertainties in the atmospheric neutrino flux.

Cut	Est. Back. Ev.	Est. Sig. Ev.	Est. Pur. [%]
0.900	311	5079	94.2
0.992	263	4864	94.9
0.994	215	4606	95.5
0.996	139	4271	96.8
0.998	118	3804	97.0
1.000	77	3017	97.5

Table 1 Estimated number of signal and background as well as the estimated purity after an application of cuts on the signalness. The number of background events was calculated rather conservative using the upper limit of the error bars.

3 Summary and Outlook

We studied the influence of a detailed feature selection using the MRMR algorithm on the training of multivariate classifiers within an atmospheric neutrino analysis for the IceCube detector. Naive Bayes, k-NN with $k = 5$ and a Random Forest were investigated.
We find that the MRMR feature selection can be considered stable if the number of attributes considered is $n_{Attributes} \geq 20$. We also studied the influence of removing correlated attributes prior to the MRMR selection on the performance of the Random Forest. We find that the most stable performance could be achieved if attributes with $\rho \geq 0.9$ are removed before running the MRMR algorithm. The optimum per-

formance was found if attributes with $\rho \geq 0.95$ were removed prior to MRMR. The difference in performance compared to $\rho \geq 0.9$, however, is negligible.
A Random Forest was trained using 500 trees and 3.4×10^5 simulated signal and 3.4×10^5 simulated background events in a 5-fold cross validation. We find that purities above 95% can be achieved depending on the signalness cut. It was shown that the number of neutrinos can exceed 3000 from ≈ 14000 at Level 3. These numbers have, however, been evaluated using Monte Carlo simulations only and might be subject to changes. The changes might be due to data MC mismatches and uncertainties in the atmospheric neutrino flux.

Acknowledgements 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

1. R. Abbasi *et al.* The Energy Spectrum of Atmospheric Neutrinos between 2 and 200 TeV with the Amanda-II Detector, Astropart. Phys. **34** (2010)
2. R. Abbasi *et al.* Measurement of the atmospheric neutrino energy spectrum from 100 GeV to 400 TeV with IceCube, Phys. Rev. D **83**, (2011)
3. M. Ackermann *et al.*, Optical properties of deep glacial ice at the South Pole, J. of Geophys. Res. **111** (2006) D13203, July 2006
4. J. Ahrens *et al.* Sensitivity of the IceCube detector to astrophysical sources of high energy muon neutrinos, Astropart. Phys. **20** (2004)
5. L. Breiman, Random Forests, Machine Learning 45 (2001)
6. T. DeYoung, Neutrino Astronomy with IceCube, Modern Physics Letters A, Vol. 24, Iss. 20 (2009)
7. C. H. Q. Ding and Hanchuan Peng, Minimum Redundancy Feature Selection from Microarray Gene Expression Data, 2nd IEEE Computer Society Bioninformatics Conference (CSB 2003) (2003)
8. D. Heck, CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers, Forschungszentrum Karlsruhe Report RZKA 6019 (1998)
9. M. Honda *et al.*, Calculation of the flux of atmospheric neutrinos, Phys. Rev. D **52**,9 (1995)
10. L.I. Kuncheva, A stability index for feature selection, Proceedings of the 25th IASTED International Multi-Conference (2007)
11. I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, YALE: Rapid Prototyping for Complex Data Mining Tasks, KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (2006)
12. E. Resconi, Status and prospects of the IceCube neutrino telescope, Nucl. Instr. and Meth. A **602**, 7 (2009)
13. B. Schowe and K. Morik, Fast-Ensembles of Minimum Redundancy Feature Selection, Workshop on Supervised and Unsupervised Ensemble Methods and their Applications - SUEMA 2010 (2010)
14. B. Schowe, http://sourceforge.net/projects/rm-featselext (2011)
15. T. Stanev, Status, performance, and first results of the IceTop array, Nucl. Phys. B (Proceedings Supplements) 196 (2004)
16. T. Hastie and R. Tibshirani and J. Friedman, The elements of Statistical Learning - Data Mining, Inference and Prediction, Springer, Berlin Heidelberg New York (2001)