Technical Report

RISE Germany Internship: Application of Data Mining Methods on IceCube Event Reconstructions

Srishti Bhasin, Mathis Börner

08/2016

technische universität dortmund

SFB 876 Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

Part of the work on this technical report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C3.

Speaker: Prof. Dr. Katharina Morik
Address: TU Dortmund University
Joseph-von-Fraunhofer-Str. 23
D-44227 Dortmund
Web: http://sfb876.tu-dortmund.de

1 Introduction

In this report the results from a 3-month internship are presented. The goal of the internship was to apply data mining methods to low level IceCube data in order to reconstruct the particle energies. IceCube [2] is a neutrino observatory located at the geographical South Pole, built with the aim of detecting high energy astrophysical neutrinos. The detector consists of 5160 photomultipliers, located 1.5-2.5 kilometers beneath the icecap, which detect Cherenkov light radiated by charged particle propagation through the ice. The reconstruction of detected events directly at the pole is challenging, due to heavy constraints on resources. Due to this, only rudimentary reconstructions [1] are performed on-site. The final results are obtained months later, once the data has been transported from the detector. An effective and prompt reconstruction directly at the pole would open a lot of new possibilities for follow-up studies of detected events. The application of state-of-the-art data mining methods can help to obtain these reconstructions on-site.

Event Detection and Estimated Energies

With the IceCube detector analyses for all flavours of neutrinos can be performed, but for this work, only muon neutrinos were examined. Due to their excellent angular resolution, follow-up studies of muon neutrino events have the highest chance to succeed. The neutrinos cannot be detected directly, but they are measured indirectly through the particles produced in their interactions with nuclei. There are two types of neutrino interactions, charged current (CC) and neutral current (NC). Both types of interaction lead to the production of a shower of hadrons, detected as a "cascade". In CC interactions of a muon neutrino, a further charged muon is produced. NC interactions, on the other hand, produce another neutrino. IceCube detects particles via their emitted Cherenkov light. In a cascade, a multitude of charged particles are created, and the signature in the detector has a spherical shape. Since a neutrino cannot be detected, the NC interactions are measured as a single cascade.

A muon can travel multiple kilometers through the ice and incurs energy losses due to bremsstahlung, pair production and ionization processes. In these processes, again, charged particles are created and their emitted Cherenkov light can be detected. This light leads to a track-like signature in IceCube. Since the neutrino interactions do not have to take place inside the detector, the most frequent signature for a detected muon neutrino is a muon track.

As already stated, the primary neutrino and its energy are not directly accessible. Its energy has to be inferred from the signature its daughter particles leave in the detector. In this work we differ between three different energies of interest: the total of energy losses inside the detector E_{Losses}, the energy of the muon produced in a CC interaction E_μ, and the energy of the primary neutrino E_ν.

The aim of this project was to estimate the energy of the primary neutrino, the energy of the produced muon, and the total energy losses inside the detector, using only attributes available on-site. In this way, one can gain an idea of how feasible it is to run real-time energy estimations at the detector itself. These estimations are carried out using

supervised machine learning techniques.

2 Method and Results

The starting point for the study are data as processed at the pole. The rudimentary reconstructions present in this data set are called "online" attributes. Machine learning based regression algorithms are applied to these attributes in order to estimated the three different energies.

2.1 Label Generation

Using MC data, one can go through each recorded event. For the E_ν the energy of the primary neutrino can be taken from the simulation. The muon energy E_μ is the energy of the muon when entering the detector. The simulation of the muon records the creation of the muon and all energy losses along its propagation. To acquire the sought-after energy, the muon must be traced up to its point of entry. For the total energy loss E_{Losses}, all energy losses occurring inside the detector are summed over. To do so, a recursive function is used to find the last neutrino interaction and identify whether it is an NC or CC interaction. In the case of an NC interaction, the sum of energy losses is the energy deposited in the hadronic cascade. For CC interactions, E_{Losses} is the sum of all energy losses of the muon and in rare occasions, the initial cascade. The simulation of the muon propagation and interaction is carried out with PROPOSAL [4].

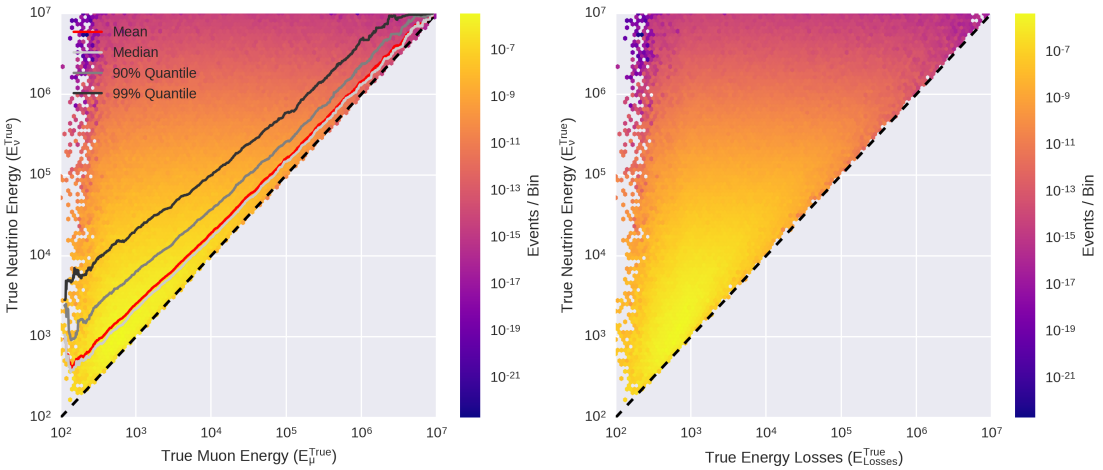
Two dimensional histograms of E_{Losses} and E_μ against E_ν are shown in Figure 1. As the plots confirm, both E_{Losses} and E_μ always have to be lower than the primary energy. The plots also show the smearing of the energies dependent on the neutrino energy, with the mode relatively close to the diagonal.

2.2 Data Mining Process

The approach carried out uses a Random Forest (RF) [3] regression for the estimation of the different energies. To fully utilize the RF algorithm, some preprocessing steps are required.

Preprocessing and Attribute Selection

The initial dataset consists of 800 000 events. Initially, events without the full online reconstructions are removed. This reduces the number of events to 460 000. Additionally, all attributes with more than 5 % NaNs are removed. Finally, the correlation of the attributes to each other is considered. Duplicate attributes were removed by considering those with very high Pearson correlation coefficients ($\rho > 0.99$). The remaining 112 attributes are used to train the RF.



(a) Muon energy E_μ with different quantiles and the mean neutrino energy.

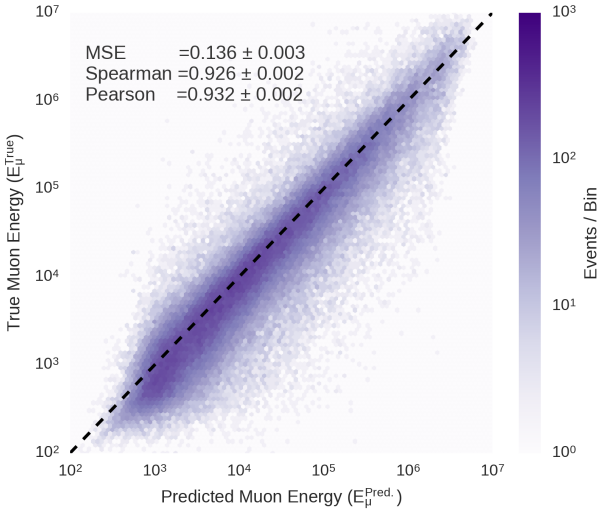
(b) Energy losses E_{Losses}

Figure 1: Two dimensional histograms of E_{Losses} and E_μ against E_ν

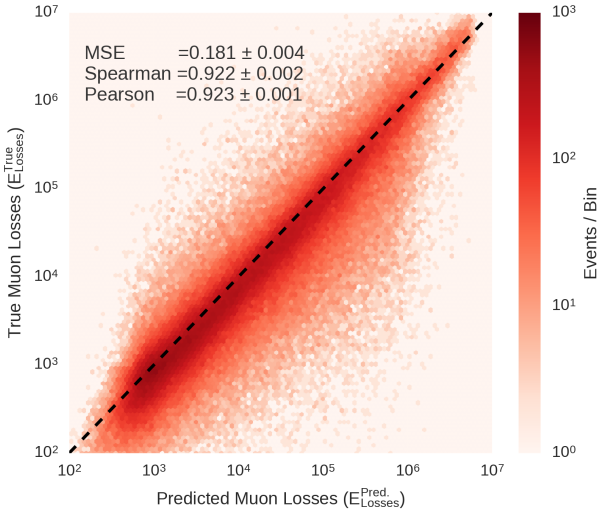
Energy Estimation

The Random Forest regression was used to estimate the labels. The RF regressor is based on decision trees, so works by finding properties that most effectively split the data using the mean squared error. It is random because every split is chosen from a random subset of the attributes, rather than all of them. Because of the robustness and the easy optimization a RF is a good starting point to get a first benchmark of the usability of machine learning algorithms for event reconstructions.

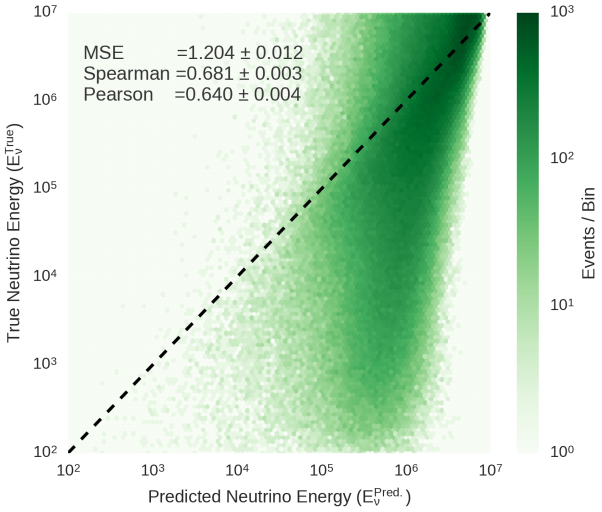
The regression was implemented with a 20-fold cross validation. In this method, the data is split into 20 roughly equally sized groups, from which 19 are used to train the algorithm and the remaining 1 is used for testing purposes. The number of estimators was set to 100. The results are displayed in Figure 2. To quantify the achieved performance, the mean squared error (MSE), the Spearman correlation coefficient, and the Pearson correlation coefficient are calculated. Each value and corresponding error was calculated by taking the mean and standard deviation of the relevant values computed for each of the 20 folds of data.



(a) Prediction of muon energy



(b) Prediction of total energy losses in detecter



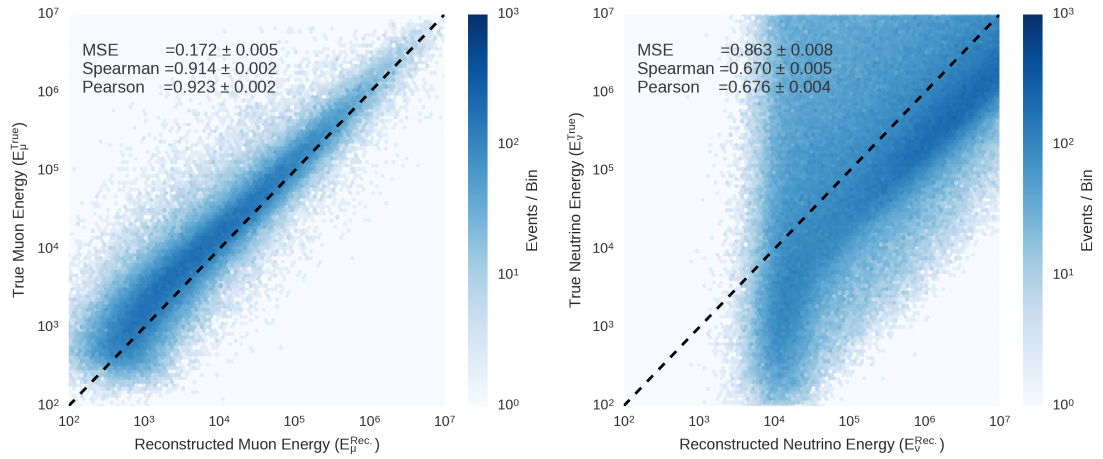
(c) Prediction of primary energy

Figure 2: Results of Random Forest regression to estimate energies using online attributes. All plots show prediction of energy against true energy. The colors of the plots in the report are used consistent: Muon Energy (purple), Energy Losses (red), Neutrino Energy (green), Reconstructed Energies (blue).

The correlation coefficients indicate that the regression was very successful in estimating the muon energy and the energy losses in the detector, see (Figure 2a and Figure 2b). However, there were limitations in the estimation of the primary energy, as is evident by the clear divergence from the ideal diagonal trend in Figure 2c. This discrepancy was partly expected, as the detector can measure the energy losses directly and infer the muon energy from those losses. However, the neutrino energy has to be inferred from the in a CC interaction produced muon. The position of the interaction vertex is unknown

and therefore the distance traveled and the amount of lost energy outside of the detector are not accessible. This makes the estimation of the neutrino energy very challenging. Nevertheless, it is worth investigating further to determine whether the fault lies with the method of regression or simply the lack of sufficient information in the online attributes, and this is discussed in subsection 2.3.

To get a better comparison for the achieved performance. The RF predictions are compared with "classical" energy estimation algorithms. The figures in 3 show the reconstruction of the muon and neutrino energies present in the online attributes. Those attributes ("truncated energy" [1]) attempt to estimate the energy of the particles using the signals measured by the DOMs in the detector. As is evident, the reconstructed muon energy performs better than the reconstructions of the primary energy.



(a) Reconstructed muon energy against true

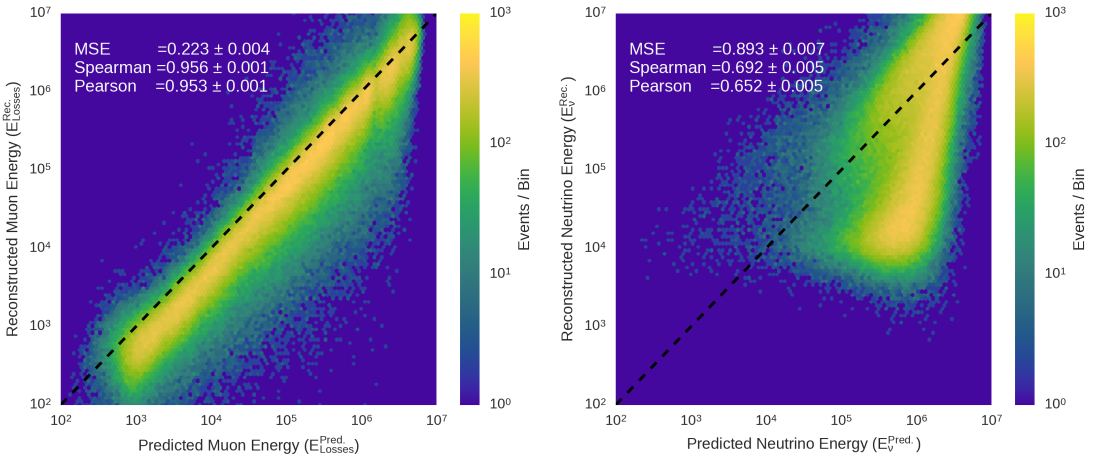
(b) Reconstructed neutrino energy against true

Figure 3: Reconstructed energies against true energies.

The comparison between RF prediction Figure 2a and the reconstruction Figure 3a for the muon energy show, that the RF prediction achieves better results and beats the reconstruction in all quantities. This is what one could expect since the RF uses the reconstruction as an input parameter. Therefore, the reconstructed energy can be taken as a lower bound on the performance of the RF. That the RF is only reproducing the reconstruction can be disproved with Figure 4a. The plot shows the correlation between the prediction and the reconstruction, and proves that the RF prediction is able to significantly improve the reconstruction of the muon energy.

For the prediction of the neutrino energy the result is quite divergent. The RF prediction achieves even worse results (see Figure 2c) in comparison to the classical reconstruction (see Figure 3b). As stated earlier, the prediction of the neutrino energy is a difficult task, since in most of the events only the produced muon which traveled an unknown distance before entering the detector is detected. The bad result of the prediction is in great contrast to the successful predictions of the muon energy. Furthermore, a RF regression does not seem to be the correct approach for the reconstruction of the incident neutrino

energy.



(a) Prediction of muon energy against reconstructed muon energy

(b) Prediction of neutrino energy against reconstructed neutrino energy

Figure 4: Results of Random Forest predictions against reconstructed energies.

2.3 Alternative estimation of primary energy

In order to improve the poor estimate of the neutrino energy, a different method of estimation was also implemented as follows. One uses the predicted muon energies and applies the smearing between the muon and the neutrino energy (see Figure 1a) on the prediction to get an estimate for the neutrino energy. Around each predicted muon energy $E^{\text{Pred.}}_{i,\mu}$, simulated events with:

$$\log\left(\left|E^{\text{True}}_{\mu} - E^{\text{Pred.}}_{i,\mu}\right|\right) < 0.1/\text{GeV}$$

are used to determine the median true neutrino energy. This median energy is the new prediction for the neutrino energy. The resulting prediction can be seen in Figure 5.

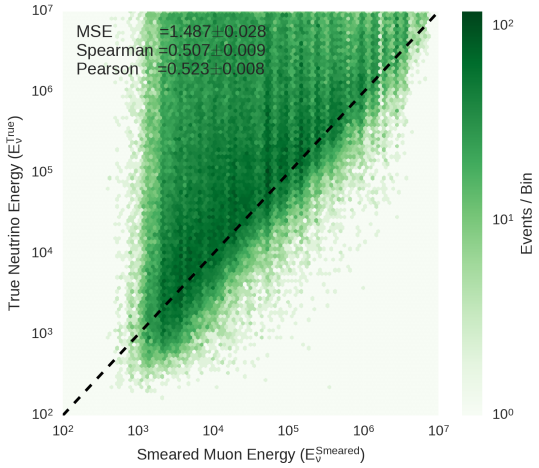


Figure 5: Smeared neutrino energy estimate against true neutrino energy.

The predicted neutrino energy tends to underestimate the true neutrino energy. In comparison to the direct predictions, the new prediction take the physical limit $E_\mu < E_\nu$ directly into account. This tendency to underestimate leads to worse correlation coefficients and a larger MSE.

Both predictions for the neutrino energy are not as good as the classical reconstructions. Moreover, the prediction from the smeared muon energy proves that it is not possible to infer the neutrino energy purely based on the muon energy. Another source of information about neutrino energy is, for example, the direction of the muon. Due to an increase of the neutrino-nucleon cross section with energy, the Earth core becomes more opaque for high energy neutrinos. The correlation between muon energy and neutrino energy for neutrinos propagating through Earth is therefore different from the connection obtained for neutrinos entering the detector from above. A multivariate approach which is better suited for low level data is expected to learn such relations, and yield better results.

3 Conclusion

The estimates of the muon energy, Figure 2a, and the energy losses, Figure 2b, show the huge potential and the direct applicability of machine learn algorithms on event reconstructions. The results for the neutrino energy showed that the chosen approaches were incapable of producing satisfactory reconstructions for neutrinos. This can be explained by the fact that the prediction of the primary energy is a very complex task, and the performance of the Random Forest algorithm is highly dependent on the input attributes.

A very promising approach would be to use state-of-art deep learning algorithms. In particular, convolutional neural networks should be tested, since they are able to produce strong input attributes themselves, and showed unmatched results in image recognition tasks. The individual signals measured by the DOMs in IceCube can be treated similar to pixels in image recognition tasks, so the use of neural networks would be an appropriate advancement.

References

[1] M. G. Aartsen et al. “Energy reconstruction methods in the IceCube neutrino telescope”. In: *Journal of Instrumentation* 9.03 (2014), P03009. URL: http://stacks.iop.org/1748-0221/9/i=03/a=P03009.

[2] A. Achterberg et al. “First year performance of the IceCube neutrino telescope”. In: *Astroparticle Physics* 26.3 (2006), pp. 155–173.

[3] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[4] J.-H. Koehne et al. “PROPOSAL: A tool for propagation of charged leptons”. In: *Computer Physics Communications* 184.9 (2013), pp. 2070–2090. ISSN: 0010-4655. DOI: http://dx.doi.org/10.1016/j.cpc.2013.04.001. URL: http://www.sciencedirect.com/science/article/pii/S0010465513001355.