Data analysis with IceCube Monte Carlo simulation data

Judith Gnade
judith.gnade@tu-dortmund.de

Hendrik Lauersdorf
hendrik.lauersdorf@tu-dortmund.de

John Wendel
john.wendel@tu-dortmund.de

April 30, 2022

TU Dortmund – Fakultät Physik

Contents

1 Theoretical foundations

1.1 Cosmic radiation

Charged cosmic radiation mainly consists of high-energy protons, helium and heavy nuclei. The exact composition depends on the energy level of the particles. Energies up to 10^{20}eV can be achieved by that kind of radiation which follows approximately the power law of

$$\frac{\mathrm{d}\Phi}{\mathrm{d}E} = \Phi_0 E^{\gamma}.$$

Here $\gamma \approx -2.7$ is the spectral index for charged particles.
The IceCube experiment studies muons and neutrinos from atmospherical and astrophysical sources. Atmospherical particles can be separated into two categories: conventional and prompt muons and neutrinos. Conventional muons and neutrinos originate from pion and kaon decays in the atmosphere. Since pions and kaons have a relatively long lifetime, they are losing a part of their energy before decaying into muons and neutrinos. This results a shift of the energy spectrum from $\propto E^{-2.7}$ to $\propto E^{-3.7}$. Besides kaons and pions there is also a production of D mesons and Λ_c baryons for high energy interactions. These particles have a very short lifetime, so they are decaying before deposing a relevant amount of energy. This results to the production of prompt muons and neutrinos with an unshifted energy spectrum of $\propto E^{-2.7}$. Astrophysical neutrinos are emitted from hadron accelerating sources that are also emitting high energy photons. These photons are not able to pass dust clouds or other optically dense media. Neutrinos on the other side have such a small cross-section that they are not interacting with these kind of boundaries. Assuming shock acceleration [3] is getting a spectral index of $\gamma \approx -2$ for astrophysical neutrinos.

1.2 The IceCube experiment

The IceCube experiment is located at the geographical southpole 1450 m-2450 m under the ice surface. Its main purpose is to detect high energy neutrinos and muons. The detector consists of 86 wires with a total of 5160 photomultipliers (PMT). These PMTs are detecting Cherenkov radiation resulting from high energy charged particles that are traveling with a velocity greater than the speed of light in that medium. The speed of light in a medium is given by

$$c = \frac{c_0}{n} \tag{1}$$

where c_0 is the vacuum speed of light and n is the refractive index.
7 wires with a bigger PMT density and a smaller distance to each other are located in the center of the detector and form the low energy expansion DeepCore [2]. The lower distances and the higher energy efficiency are resulting in a relatively low energy threshold of 10 GeV. The rest of the wires that surround DeepCore are forming the In-Ice-Array, the biggest part of the detector. Since the distances between the PMTs are bigger than in DeepCore, the energy threshold has a value of 100 GeV. The air shower experiment IceTop is located on the surface [1]. The Cherenkov radiation on this part of

the experiment is measured in ice tanks. IceTop serves to study cosmic radiation as well as a veto for the In-Ice-Array. A schematic representation of the IceCube experiment is shown in figure 1.

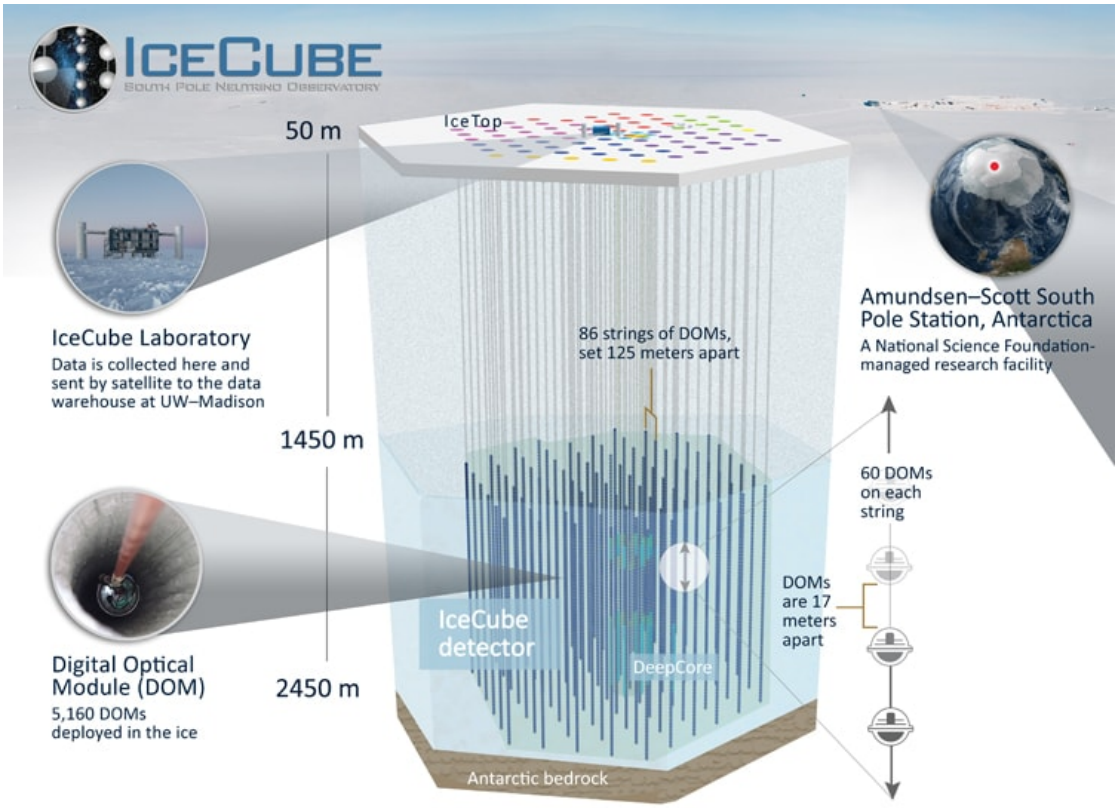

Figure 1: Schematic structure of the IceCube experiment [4].

Neutrinos are measured via secondary particles of the charged current

$$\nu_l(\bar{\nu}_l) + A \to l^{\mp} + X$$

or the neutral current

$$\nu_l + A \to \nu_l + X\,.$$

Due to their high energy loss electrons have a spherical signature of Cherenkov light while muons have a longer signature since their energy loss is much lower. Tau leptons possess a similar spherical signature as electrons since the life time is very low resulting in a low reach before decaying. In the neutral current a cascade induced by the secondary hadrons is observed which has also a similar signature as electrons. Cherenkov light from muons is too weak to be detected but muons interacting with the medium are producing e^+e^- pairs and photons which are resulting in a cascade. The Cherenkov light of these secondary particles can be measured by the experiment.

1.3 Measurement of neutrinos with the IceCube experiment

An analysis method to reject atmospherical muons is to use *starting events*. In this technique the outer layers of the detector are used as a veto for these atmospherical muons. All neutrino flavours are taken into account equally. Therefore the biggest contributions of events come from the neutral current and electron neutrino and tau neutrino via the charged current. These events have a good energy and a poor angular resolution. Tracks of muons on the other hand have a poor energy and a good angular resolution as well as a bigger reach. Planet earth itself can be used as a shield for atmospheric muons since they would be absorbed mostly. In result muons that come from that side have to be from neutrino interactions. To separate atmospherical muons from muon neutrinos a cut on the zenith angle can be made. Since the track reconstruction is not perfect a cut on the zenith angle is just resulting in an improvement of the signal to background ratio from $1:10^6$ to $1:10^3$. To improve the separation from muons and muon neutrinos even further procedures of machine learning are used. The separation of signal and background is the aim of this analysis.

2 Multivariate selection and machine learning

In this analysis three different machine learning algorithms are used to improve the separation of signal and background. In the following these algorithms are discussed as well as the quality parameter and cross validation.

2.1 Naive-Bayes learner

The Naive-Bayes theorem is based on Bayes' theorem:

$$p\left(A|B\right) = \frac{p\left(B|A\right)p\left(A\right)}{p\left(B\right)}\,,$$

in which B is an attribute and A is the class affiliation with A representing signal and $\bar{A}$ representing background. This classifier is based on the naive assumption that the attributes are independent. If more than one attribute is looked at the quantity

$$Q = \prod_{i=1}^{n} \frac{p\left(B_i|A\right)}{p\left(B_i|\bar{A}\right)}$$

is giving a value of the probability of an event being signal like ($Q > 1$) or background like ($Q < 1$).

2.2 Random Forest Classifier

The *RandomForestClassifier* is based on binary decision trees. A decision tree is dividing the data set in each step in one attribute on a specific value. The data subsets are again divided into more subsets until a defined number of steps is reached or until all subsets only contain one class (signal or background) each. To minimize effects of overtraining,

where the classifier is trained on statistical fluctuations, the average value of decision trees is taken. These trees are trained on different random subsets of the dataset.

2.3 kNN Classifier

The *k-Next-Neighbour Classifier* (kNN) is a relatively simple classifier. By looking at different distances (for example euclidean distances) the k next neighbours of an event are classified. This learning algorithm is very dependent on the choice of k since a low value is sensitive to statistical fluctuations and a high value could have too many events of different classes in one bunch. The dependency of the value of k is shown in figure 2.

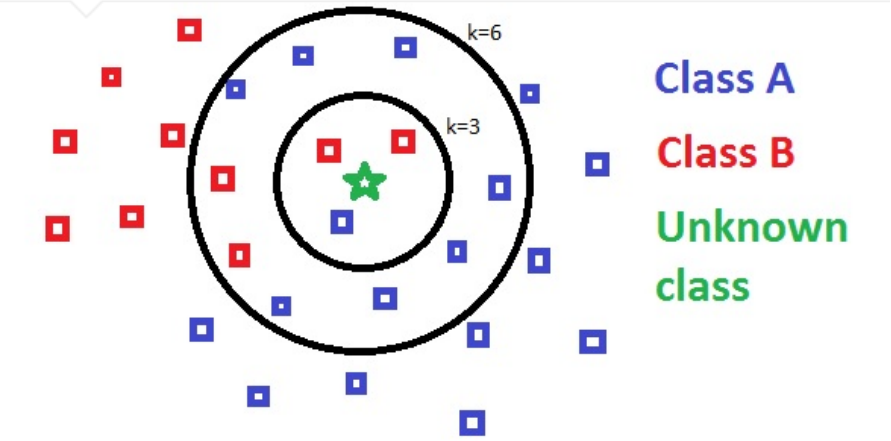

Figure 2: Example how the choice of k changes the classification of an event. With $k = 3$ the unknown event is classified to class B and with $k = 6$ it is classified as class A [5].

2.4 Quality Parameters

The quality of the separating power of a classifier is evaluated by different quality parameters. The first two parameters are the purity and the efficiency of the data set:

$$\text{purity } p = \frac{tp}{tp + fp}$$
$$\text{efficiency } r = \frac{tp}{tp + fn}\,.$$

The value tp (true-positive) means that a signal event was correctly classified as a signal event while fp (false-positive) means that a background event was falsely classified as a signal event. Similarly fn (false-negative) means that a signal event was falsely classified as a background event.

The Jaccard index is a value describing how stable the chosen attributes are against statistical fluctuations. The value is calculated like this:

$$J(F_a, F_b) = \frac{|F_a \cup F_b|}{|F_a \cap F_b|}$$

This is describing how similar the two sets F_a and F_b are. To describe how stable a dataset is the following calculation is used:

$$\hat{J} = \frac{2}{l\,(l-1)} \sum_{i=1}^{l} \sum_{j=i+1}^{l} J\big(F_i, F_j\big) \ .$$

It means that the attribute selection is been used l times on l subsets of the dataset. A value near 1 means that the dataset is stable for statistical fluctuation.

2.5 Crossvalidation

To check if a classifier did not implement statistical fluctuations the dataset is divided into n subsets. The classifier is trained on $n-1$ of these subsets and then tested on the remaining subset. This method will be repeated n times so that every subset is used as a test subset once. The resulting n values for purity and efficiency are taken into average.

2.6 ROC curves

A reciever operator curve (ROC curve) shows the ratio of *true positives* against *false positives*. A straight line between $(0, 0)$ and $(1, 1)$ represents a model with total random classification since for every value of the classification the same number of true positive and false positive events is given. A perfect classification would be shown in a curve that goes straight up to the top left corner and then to the top right corner. According to that the area under the curve (ROC AUC score) is a metric that shows the separation power of a classifier. A ROC AUC of 0.5 corresponds to a total random classification while a ROC AUC near 1 is showing a good separation power.

3 Analysis

3.1 Data preparation

Before starting the analysis the given data is being prepared. Two Monte Carlo data sets are given, one background data set and one containing signal. Every attribute that is contained in only one of the datasets is being removed as well as Monte Carlo truths, weights and event identification numbers. In the remaining attributes events that contain the value "inf" or "NaN" are being removed as well. To differentiate background and signal later labels are being created for the remaining data. Signal events are being labeled with "1" and background events are being labeled with "0".

3.2 Feature selection

The number of the features and the actual features have to be selected before getting into the classifiers. For that the Jaccard index is calculated depending on the number of features after using the *KNeighborsClassifier*. This results to a maximum Jaccard index at 20 features. For the rest of the analysis the best 20 features are determined by the *SelectKBest* method. The quality of these features is determined with `f_classif`. The data set is being split into test and training subsets to apply the classifier on.

3.3 Classifiers

Three different classifiers are used to separate background from signal: The `Naive-Bayes` Classifier, the `RandomForestClassifier` and the `KNeighborsClassifier` from the `sklearn` package are being used.
For the `RandomForestClassifier` a `RandomForest` with 100 decision trees is being created. For the classification efficiency, purity and the Jaccard index are being calculated.
For the `KNeighborsClassifier` the amount of neighbors is set to 20. It shows that a higher amount of neighbors is not increasing the separating power.
Lastly the `Naive-Bayes` Classifier is tested on the data set.
All results are shown in table 1. The different ROC curves are shown in 3.

Classifier	Efficiency	Purity	Jaccard index
RandomForest	$0{,}9771 \pm 0{,}0057$	$0{,}9893 \pm 0{,}0033$	$0{,}9669 \pm 0{,}0057$
KNeighborsClassifier	$0{,}8384 \pm 0{,}0080$	$0{,}8034 \pm 0{,}0068$	$0{,}6957 \pm 0{,}0070$
Naive-Bayes	$0{,}7886 \pm 0{,}0489$	$0{,}8036 \pm 0{,}0430$	$0{,}6605 \pm 0{,}0256$

Table 1: Efficiency, purity and Jaccard index of the three classifiers applied to the data set.

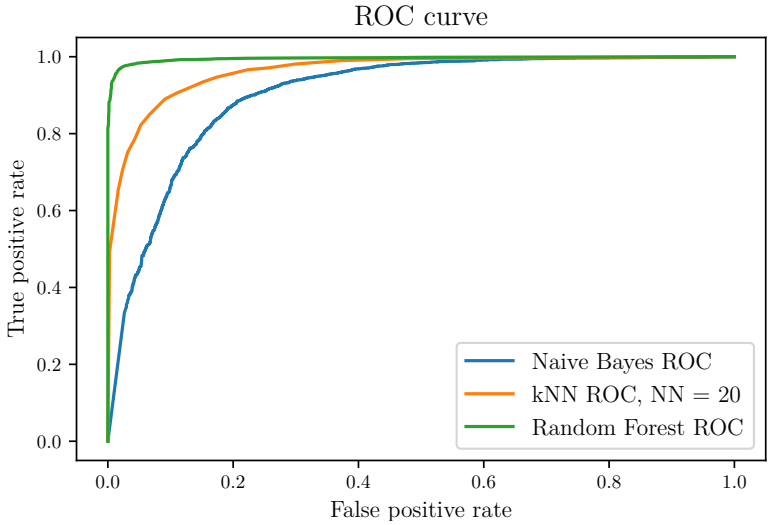

Figure 3: ROC-curves of all tested classifiers.

4 Discussion

The Naive-Bayes learner is performing the worst out of the three learners. It is the fastest of the three classifiers at the cost of separation power. The ROC AUC is the lowest as well as the Jaccard score meaning it is not as stable to statistical fluctuation. With the lowest efficiency of 0.7886 ± 0.0489 and a purity of 0.8036 ± 0.0430 it has the highest uncertainties of the three algorithms as well. That is why the Naive-Bayes learner is not suitable for this kind of separation.
The kNN classifier is showing some improvement compared to the previous learner. It is relatively fast if k is set to a low amount but for $k = 20$ the algorithm works pretty slow. If that value is being increased the classifier is even slower without any improvement of the ROC AUC. With a slightly higher efficiency of 0.8384 ± 0.0080 and a similar purity of 0.8034 ± 0.0068 it performs a bit better than Naive-Bayes. Noticable is the much lower uncertainty compared to the Naive-Bayes learner.
The `RandomForestClassifier` is the best classifier on this data set. It possesses the highest ROC AUC as well as the highest efficiency (0.9771 ± 0.0057) and highest purity (0.9893 ± 0.0033). This is because the classifier decorrelates the decision trees to search for correlations between the features. Also it is very stable to statistical fluctuation due to the fact that the classifier is trained multiple times and the average of the classifications of these trees is taken. This classifier has the best separation power according to that but it also seems to be the slowest. Deeper studies on the selected features could improve the performance but even without that this classifier seems highly qualified to separate background and signal since it has a very good ROC AUC.

References

[1] R. Abbasi et al. "IceTop: The surface component of IceCube." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 700 (2013), pp. 188–220. ISSN: 0168-9002.

[2] R. Abbasi et al. "The design and performance of IceCube DeepCore." In: *Astroparticle Physics* 35.10 (2012), pp. 615–624. ISSN: 0927-6505.

[3] E. Fermi. "On the Origin of the Cosmic Radiation." In: *Phys. Rev.* 75 (8 1949), pp. 1169–1174.

[4] *IceCube*. URL: `https://icecube.wisc.edu/science/icecube/` (visited on 05/28/2021).

[5] M. Sanjay. *KNN using scikit-learn*. URL: `https://laptrinhx.com/knn-using-scikit-learn-392440878/` (visited on 05/29/2021).