

**Arbeit zur Erlangung des akademischen Grades
Bachelor of Science**

**IceCube DeepCore Niederenergie
Veto-Studie**

Philipp Hoffmann
geboren in Datteln

2016

Lehrstuhl für Experimentelle Physik V
Fakultät Physik
Technische Universität Dortmund

Erstgutachter: Prof. Dr. Dr. Wolfgang Rhode
Zweitgutachter: Prof. Dr. Kevin Kröninger
Abgabedatum: 12. September 2016

Kurzfassung

In dieser Arbeit wird ein Vergleich zweier Vetoregionen für den IceCube Detektor im Energiebereich von 1 GeV – 1000 GeV durchgeführt. Die Vetoregionen (DC und EXT) unterscheiden sich in ihrer Grundfläche. Auf Basis von simulierten Ereignissen atmosphärischer Myonen, Elektron- und Myon-Neutrinos werden durch Anwenden der Vtos zwei Datensätze generiert. Nach vorheriger Attributsselektion durch mRMRe werden diese in einer zweistufigen Separation mit Random Forests separiert. Es werden zuerst alle Neutrinoereignisse von Myonereignissen getrennt. Von den übrigen Neutrinoereignissen werden Myon-Neutrinos als Signal selektiert. Nach jeder Separation werden die relativen Differenzen der Ereignisraten der beiden Vtos und die Korrelation zwischen rekonstruierter und wahrer Energie der Ereignisse verglichen. Mit einer erwarteten Myon-Neutrino-Rate von 4.63 μHz und einer Korrelation von 0.7 zeigt sich das DC-Veto dem EXT-Veto überlegen.

Abstract

Within this thesis a comparison between two veto regions (DC and EXT) for the IceCube detector is made in the 1 GeV – 1000 GeV energy range. Simulation data consisting of muons, electron neutrinos and muon neutrinos is used. Applying the vetoes to those events yields two new sets of events, each consisting only of events passing the corresponding veto. Relevant features are selected via mRMRe and used in a two staged random forest separation. First, all neutrinos are separated from muons, then muon neutrino candidates are separated from the remaining electron neutrinos. For each stage the passing event rates and correlation between reconstructed and true energy are compared. The DC-Veto is shown to be superior, with an expected muon neutrino rate of 4.63 μHz and a correlation of 0.7.

Inhaltsverzeichnis

1 Einleitung	1
2 Grundlagen	3
2.1 Astroteilchenphysik	3
2.2 IceCube	4
2.3 Maschinelles Lernen	8
3 Analyse	11
3.1 Neutrinoelektion	12
3.2 Myon-Neutrinoelektion	16
3.3 Vergleich der Vetoregionen	18
4 Zusammenfassung und Ausblick	25
A Anhang	27
A.1 Verwendete Software	27
A.2 NaN-Verteilung	27
A.3 Vergleich von gewichteten und ungewichteten Korrelationskoeffizienten	28
A.4 Vergleich von mRMR Methoden anhand von Testseparationen	29
A.5 Separationsqualität der EXT-Veto Neutrinoelektion	31
Literatur	35

1 Einleitung

In der Astroteilchenphysik werden entfernte Objekte im Weltall mit Teleskopen beobachtet, um über ausgesandte Teilchen, der so genannten kosmischen Strahlung, Informationen zu gewinnen. Die kosmische Strahlung besteht aus Photonen, sowie geladenen und ungeladenen Teilchen. Mit den gewonnenen Informationen ist es möglich Rückschlüsse auf ihre Quellen und Produktionsmechanismen zu ziehen. Neutrinos sind ein Teil der ungeladenen kosmischen Strahlung und können mit dem IceCube [Ach+06] Teleskop am Südpol detektiert werden. Da sie nur schwach wechselwirken werden sie indirekt durch das Tscherenkovlicht von dabei entstehenden geladenen Teilchen gemessen. Der geladene Teil der kosmischen Strahlung produziert bei Interaktionen in der Atmosphäre Neutrinos und kann somit indirekt untersucht werden. Von diesen so genannten atmosphärischen Neutrinos sind Myon-Neutrinos die zahlreichsten. In dieser Analyse sind sie Ziel der Untersuchungen und werden fortan als Signal bezeichnet. Auf Grund des geringen Wirkungsquerschnitts von Neutrinos mit Materie sind diese Signalereignisse im Vergleich zu anderen Ereignissen wie atmosphärischen Myonen, dem dominierenden Untergrund, sehr selten. Um einen hoch reinen Datensatz an Signalereignissen zu erhalten, ist es notwendig diese vom Untergrund zu separieren. Eine wichtige Vorstufe dafür sind Vtos. Mit ortsbasierten Vtos werden Teilchen verworfen, deren Wechselwirkung wahrscheinlich nicht im Detektionsvolumen stattgefunden hat. Diese Entscheidung erfolgt anhand des geschätzten Ladungsschwerpunkts und weiteren Informationen der Ereignisse. Vor allem im betrachteten Energiebereich, dem Niederenergiebereich von IceCube, sind solche Vtos besonders wichtig. Nach theoretischen Flussmodellen sind die Neutrinozahlen hier im Verhältnis zu den Untergrundzahlen besonders klein. DeepCore eignet sich besonders zur Messung dieser Neutrinos, da er dichter instrumentiert ist und diese Ereignisse genauer auflösen kann.

Ist die Wahl des Detektionsvolumens größer als DeepCore ist auf Grund der geringeren Dichte an Sensoren zu erwarten, dass die rekonstruierte Energie der Ereignisse weniger stark mit ihrer wahren Energie korreliert. Es ist auch zu erwarten, dass bei einem größeren Vetovolumen mehr Untergrundereignisse verworfen werden, während die Rate an detektierten Signalereignissen kleiner wird. In Hinsicht auf diese wesentlichen Unterschiede werden zwei Vetoregionen DC und EXT, von denen EXT das größere Detektionsvolumen hat, verglichen. Dies geschieht auf Basis von simulierten Myon-, sowie Myon-Neutrino- und Elektron-Neutrino-Ereignissen. Da auf

1 Einleitung

experimentellen Daten der Ereignistyp nicht bekannt ist, müssen zunächst Myon-Neutrinos von den restlichen Ereignissen separiert werden. Von den separierten Ereignissen werden die erwähnten Korrelationen und Raten verglichen.

2 Grundlagen

In den folgenden Abschnitten werden die Grundlagen der Astroteilchenphysik, der kosmischen Strahlung, der Neutrinointeraktionen, des IceCube Detektors und der untersuchten Vtos beschrieben. Außerdem werden die in der Analyse verwendeten Verfahren des maschinellen Lernens dargelegt.

2.1 Astroteilchenphysik

Die Astroteilchenphysik beschäftigt sich unter anderem mit der Untersuchung von kosmischer Strahlung (CR). Als Quellen der CRs werden Objekte wie Aktive Galaktische Kerne oder Gammablitze angenommen. Das Verhalten dieser Quellen kann anhand der Eigenschaften ihrer ausgesandten CRs bestimmt werden. In ihnen wird beispielsweise durch Akkretion von Materie potentielle Gravitationsenergie, und über Drehimpulserhaltung auch Rotationsenergie, in kinetische Energie der CRs umgewandelt. Ein Ansatz für die Produktion der CRs ist die Fermibeschleunigung, die durch Schocks mit interstellarer Materie Teilchen beschleunigt. Diese Art der Beschleunigung sorgt für einen CR Fluss der Form $\Phi = dN/dE$. Dieser folgt einem Potenzgesetz der Form $\Phi \propto E^{-\gamma}$ [Fer49], mit der Teilchenenergie E und dem spektralen Index γ . Die CRs können in drei Kategorien eingeteilt werden: Photonen, ungeladene Teilchen und geladene Teilchen. Photonen zeigen direkt auf ihre Quelle. Sie werden von Objekten wie Gaswolken und durch Wechselwirkung mit Sternenlicht unter Paarproduktion von Elektronen absorbiert, was einen Nachweis erschwert. Ungeladene Teilchen, wie Neutrinos oder Neutronen, werden ebenfalls nicht abgelenkt. Geladene Teilchen, wie Protonen, Elektronen oder Pionen, werden durch Magnetfelder, deren Stärke und räumliche Verteilung im Allgemeinen nicht bekannt ist, abgelenkt. Sie werden außerdem mit einer hohen Wahrscheinlichkeit von Materie absorbiert. Dadurch wird für sie sowohl der Nachweis als auch die Rekonstruktion der Quellposition erschwert. Die Hauptbestandteile der geladenen CRs sind Protonen und Heliumkerne mit Anteilen von 90% und 9% [GER16]. [Cou13]

2 Grundlagen

Gelangen die Protonen bis zur Erde, können sie mit Atomkernen N in der Atmosphäre interagieren und Mesonen, sowie weitere Interaktionsprodukte X erzeugen:

$$p + N \rightarrow \begin{cases} \pi^\pm + X \\ K^\pm + X \end{cases} \quad (2.1)$$

Die Pionen und Kaonen zerfallen wiederum über:

$$\pi^\pm \rightarrow \mu^\pm + \nu_\mu (\bar{\nu}_\mu) \quad (\sim 100\%) \quad (2.2)$$

$$K^\pm \rightarrow \mu^\pm + \nu_\mu (\bar{\nu}_\mu) \quad (\sim 63.5\%) \quad (2.3)$$

$$\mu^\pm \rightarrow e^\pm + \nu_e (\bar{\nu}_e) + \bar{\nu}_\mu (\nu_\mu) \quad (2.4)$$

Produkte solcher Interaktionen in der Atmosphäre werden auch *atmosphärisch* genannt. Wird von atmosphärischen Teilchen gesprochen, bezieht sich dies auf Interaktionsprodukte, die in der Atmosphäre entstehen. Mit ihnen sind Rückschlüsse auf die Eigenschaften der primären CR möglich. Die Neutrinos aus den Reaktionen (2.2) bis (2.4) haben durchschnittlich eine etwa gleich große kinetische Energie, die etwa einem Zehntel der Energie des Protons entspricht. Für die im Folgenden betrachteten atmosphärischen Neutrinos im Energiebereich von 100 GeV – 1000 GeV wird ein Spektrum mit $\gamma = 2.67$ erwartet. [Oli+14; WBM98; GER16]

Neutrinos interagieren nur über die schwache Wechselwirkung, womit sie nicht auf direktem Weg nachgewiesen werden können. Werden bei solch einer Interaktion geladene Leptonen erzeugt, ist es möglich diese über ihr Tscherenkowlicht in einem transparenten Medium nachzuweisen. Damit dieses Licht erzeugt werden kann, müssen sich die Sekundärteilchen mit einer Geschwindigkeit größer der Lichtgeschwindigkeit im Medium bewegen. Der für diesen Prozess wichtigste Reaktionstyp läuft über den geladenen Strom (CC) der schwachen Wechselwirkung ab.

$$\nu_\ell (\bar{\nu}_\ell) + N \rightarrow \ell^- (\ell^+) + X \quad (\text{CC}) \quad (2.5)$$

Hierbei interagiert ein Neutrino ν des Flavours ℓ mit einem Atomkern N zu einem Lepton ℓ des gleichen Flavours und einer hadronisierten Kaskade X . Die Anzahl so interagierender Neutrinos pro Jahr wird für einen 1 km^3 großen Detektor wie IceCube im Energiebereich von 1 GeV – 1000 GeV auf über 100 000 geschätzt.

2.2 IceCube

IceCube [Ach+06] ist ein Teleskop für hochenergetische Neutrinos am geografischen Südpol. Es besteht aus dem auf antarktischen Eis gelegenem IceTop und dem im

Eis eingelassenen IceCube In-Ice Array. IceTop wird unter anderem als Veto für atmosphärische Myonen verwendet, Details dazu sind in [Abb+13] zu finden. Das IceCube In-Ice Array besteht aus 80 so genannten Strings, an denen insgesamt 4800 Digitale Optische Module (DOMs) angebracht sind. Die Strings sind mit einem mittleren Abstand von 125 m untereinander in das Eis eingeschmolzen. Die gleichmäßig auf jedem String verteilten DOMs befinden sich in einer Tiefe von 1450 m bis 2450 m und dienen als Detektoren von Tscherenkowlicht. Es werden also nur von Neutrinos oder anderen Teilchen induzierte Ereignisse gemessen. Diese werden im Folgenden nur Neutrinos, Neutrinoereignisse, etc. genannt. Die Detektion geschieht durch eine Photomultiplierröhre, dessen Ausgang mit einem Data Acquisition Board verbunden ist, das die gemessenen Spannungspulse zum IceCube-Labor an der Oberfläche weiterleitet. Die Konstruktion ist von einer Glaskugel umgeben und wird so vor dem Druck des Eises geschützt. Näheres zu den DOMs ist in [Abb+09] zu finden.

DeepCore

DeepCore (DC) ist ein dichter instrumentiertes Teilarray von IceCube, das zur Detektion von niederenergetischen Neutrinos verwendet wird. DC besteht aus der innersten Stringsschicht des In-Ice-Arrays sowie sechs weiteren DC-Strings, die zusammen einen mittleren Abstand von nur 72 m haben. Die genaue String-Konfiguration ist in Abbildung 2.1 dargestellt. Die DC-Strings sind mit HQE-DOMs, die eine höhere Quanteneffizienz als die im restlichen Array verwendeten DOMs haben, ausgestattet. Der DOM-zu-DOM-Abstand an einem String liegt zwischen 7 m und 10 m, wobei im Bereich der Staubschicht von –2000 m bis –2100 m keine DOMs liegen. Die Anzahl an DOMs pro Volumen ist in DC etwa fünf mal so hoch wie im IceCube In-Ice Array. All dies sorgt dafür, dass in DC bei 10 GeV Teilchenenergie bereits etwa zehn DOMs ansprechen und die optimale Detektionsenergie im Bereich von 10 GeV – 100 GeV liegt. Weitere Informationen sind in [Abb+12] zu finden.

Vetoregionen

Um bei Messungen von – für IceCube – niederenergetischen Neutrinos im Energiebereich von 10 GeV – 100 GeV eine ausreichende Energieauflösung zu erreichen, wird ein Detektor mit einer durchschnittlich höheren DOM-Dichte als im IceCube In-Ice Array benötigt. Deswegen sollte für solche Ereignisse nicht das ganze IceCube-Array zur Detektion benutzt werden, sondern nur DeepCore und gegebenenfalls noch eine begrenzte Anzahl an umliegenden String-Schichten. Die ungenutzten Strings können als Vetoregion genutzt werden, um das hohe Neutrino zu Myon Verhältnis von $1:10^6$

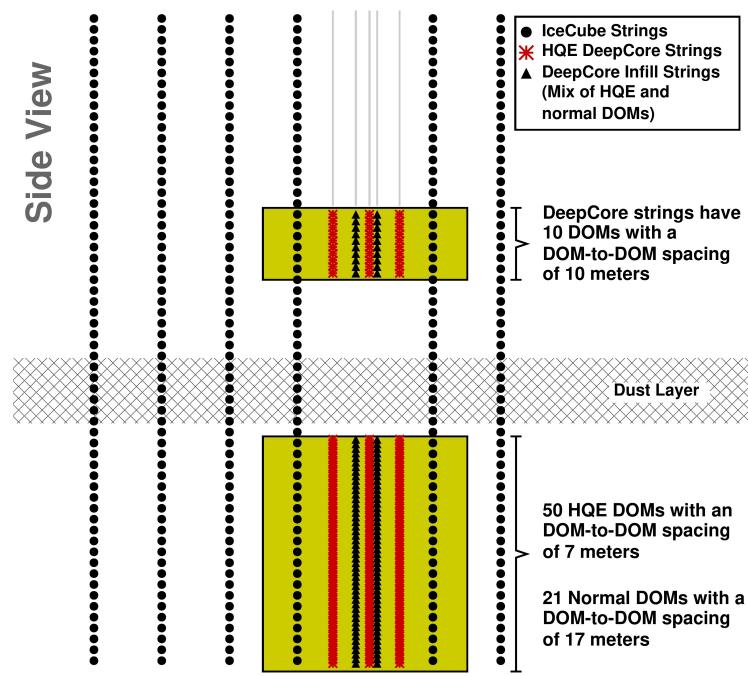


Abbildung 2.1: Eine Seitenansicht der String- und DOM-Konfiguration des IceCube-Detektors. Der herkömmliche DeepCore Bereich ist grün unterlegt. Weiter gekennzeichnet sind die HQE-DOMs und die Lage der Staubschicht.

zu verringern. Für das Veto werden Ereignisse betrachtet, die innerhalb DeepCores und somit außerhalb der Vetoregion starten. Beim Filtern von Ereignissen durch ein Veto werden verschiedene Kriterien berechnet. Wird eines der Hauptkriterien oder eine gewisse Kombination von Nebenkriterien erfüllt, wird ein Ereignis vom Veto verworfen. Für die übrig bleibenden Ereignisse liegt die Wahrscheinlichkeit sehr hoch, dass sie im zum Veto gehörenden Detektionsvolumen gestartet sind. Eines der Hauptkriterien ist die Geschwindigkeit des Ereignisses. Berechnet wird sie aus der Distanz und vergangenen Zeit zwischen Startpunkt und dem Ladungsschwerpunkt des Ereignisses. Ist sie im Bereich der Lichtgeschwindigkeit wird das Kriterium erfüllt und das Ereignis verworfen. Näheres dazu ist in [Dau+12] zu finden.

In Abbildung 2.2 sind die Bereichsdefinitionen der in dieser Arbeit betrachteten Vatos dargestellt. Beim sogenannten DC-Veto wird das DC Array und eine zusätzliche Schicht an Strings zum Detektionsvolumen gezählt. Das Detektionsvolumen des EXT-Vetos besteht aus dem DC-Array und zwei weiteren Schichten an Strings.

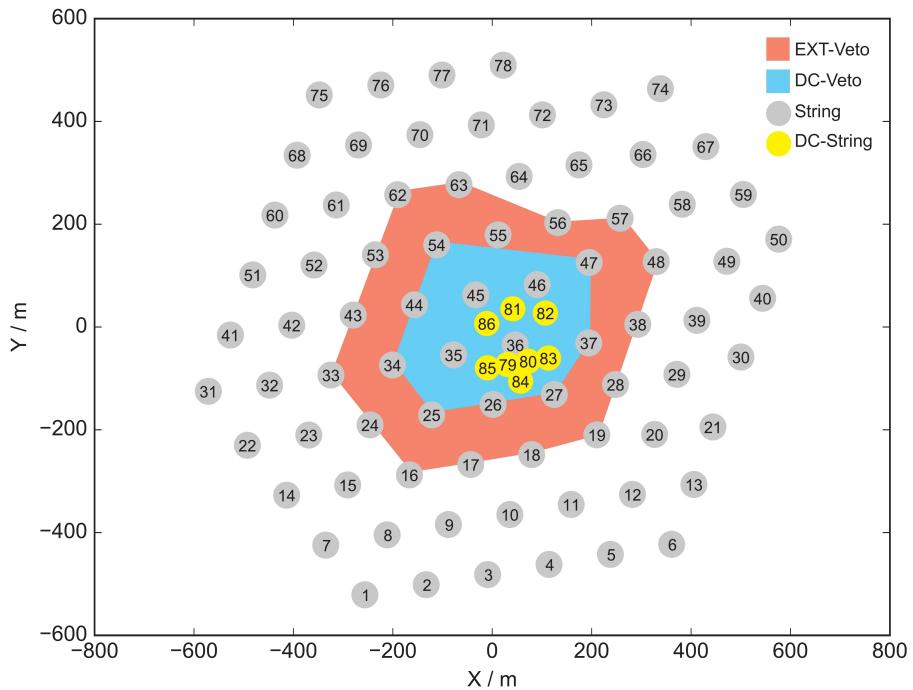


Abbildung 2.2: Eine Draufsicht der IceCube String-Konfiguration. Das DC-Detektionsvolumen ist blau unterlegt. Rot unterlegt ist das erweiterte Detektionsvolumen des EXT-Vetos. Strings außerhalb des Detektionsvolumens zählen zur Vetoregion.

2.3 Maschinelles Lernen

Die detektierten Ereignisse sind anfangs noch ungetrennt. Um beispielsweise die Energiespektren der verschiedenen Teilchensorten zu bestimmen ist jeweils ein hoch reines Sample erforderlich. Deshalb müssen die Ereignisse voneinander getrennt werden. Die Trennung der Ereignisse in die beiden Klassen Signal und Untergrund ist durch hohe Datenraten von mehr als 100 GB pro Tag per Hand unmöglich. Abhilfe schaffen Methoden des maschinellen Lernens. Beim maschinellen Lernen wird ein Algorithmus, weiter Lerner genannt, auf die Daten angewendet. Als Eingabe bekommt der Lerner einen Datensatz aus Ereignissen mit Attributen, die diese Ereignisse charakterisieren. Durch Veränderung von Modellparametern, die das Verhalten des Lerners bestimmen, löst dieser ein Optimierungsproblem; er wird auf den Daten trainiert. Haben die Eingabedaten die zusätzliche Information der Klassenzugehörigkeit, auch Label genannt, fällt das Verfahren in die Kategorie des überwachten Lernens. In der später durchgeführten Analyse werden hierfür Simulationsdaten benutzt. Der so trainierte Lerner kann eingesetzt werden, um für Daten des selben Typs eine Vorhersage über ihre Klassenzugehörigkeit zu treffen. Dieser Schritt wird auch Separation genannt. Die in diesem Fall zu trennenden Ereignisse fallen in zwei Kategorien: Weiter als Signal und Untergrund bezeichnet. Es gibt auch multiklassen Probleme. [Agg15]

Attributsselektion mittels mRMR

Eine Attributsselektion kann hilfreich sein, um den Rechenaufwand des in der Separation verwendeten Lerners zu verringern. Dafür kann beispielsweise ein Algorithmus namens Minimum Redundancy Maximum Relevance (mRMR) verwendet werden. Dieser basiert auf der gegenseitigen Information (MI) der Attribute untereinander, die aus den unabhängigen Wahrscheinlichkeitsdichten $p(x)$ und $p(y)$, sowie der gegenseitigen Dichte $p(x, y)$ über

$$MI(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad (2.6)$$

berechnet wird. x und y sind Ausprägungen beliebiger Attribute X und Y . Mit der MI lassen sich die Parameter Relevanz D und Redundanz R berechnen. Die Relevanz eines Attributs ergibt sich aus der MI des Attributs mit dem Zielattribut. Die Redundanz eines Attributs ergibt sich aus der MI des Attributs und allen anderen Attributen. Die für kontinuierliche Daten notwendigen Anpassungen von MI, D und R sind in [DP05] dargelegt. Die nach der Selektion übrig bleibende Menge von k Attributen wird über Maximierung eines Optimierungskriteriums $\Phi(R, S)$ gefunden.

Für Φ wird meist die Differenz der gegenseitigen Information (MID) $\Phi = D - R$, oder der Quotient der gegenseitigen Information (MIQ) $\Phi = D/R$ verwendet.

Das in der eigentlichen Separationskette verwendete Paket mRMRe [De +13] berechnet die MI über eine lineare Näherung mit dem Pearson-Korrelationskoeffizienten ρ zu $MI = -\frac{1}{2} \ln(1 - \rho(X, Y))^2$. Als Optimierungskriterium wird eine Art der MID verwendet, die die durchschnittliche Redundanz zu bereits ausgewählten Attributen minimiert. Vorteil dieser Methode ist eine kürzere Laufzeit.

Um eine Aussage über die Stabilität der Selektion treffen zu können, wird die Attributselektion mehrmals auf disjunkten Teilmengen von Ereignissen wiederholt. Als Maß für den Unterschied zweier Mengen ausgewählter Attribute F_a und F_b dient der Jaccard-Index J [Jac12]:

$$J = \frac{|F_a \cap F_b|}{|F_a \cup F_b|} \quad (2.7)$$

Werden n Selektionen für den Stabilitätstest durchgeführt, können alle Jaccard-Indices für die möglichen Mengenkombinationen zu einer Gesamtstabilität gemittelt werden:

$$\hat{J} = \binom{n}{2}^{-1} \sum_{i=1}^n \sum_{j=i+1}^n J(F_i, F_j) \quad (2.8)$$

Klassifizierung mittels Random Forest

Der Random Forest dient zur Klassifizierung von Ereignissen und wird auf gelabelten Daten, solche für die die Klassenzugehörigkeit bekannt ist, trainiert. Mit dem trainierten Random Forest ist es anschließend möglich eine Vorhersage der Klassenzugehörigkeit von ungelabelten Daten zu treffen. Der Random Forest besteht aus mehreren Entscheidungsbäumen [Qui86]. Jedem Baum steht eine zufällig gezogene Teilmenge an Ereignissen aus der Gesamtmenge an Ereignissen zur Verfügung. Auf dieser Teilmenge werden schrittweise Schnitte durchgeführt, anhand derer ein Ereignis als Signal oder Untergrund klassifiziert wird. Jeder Schnitt wird so gewählt, dass er ein Optimierungskriterium maximiert. Für ein Ereignis erzeugt jeder Baum so eine Zuordnung. Bei einem Zweiklassenproblem lassen sich die Zuordnungen eines Baums als Zahlenwerte 0 (Untergrund) und 1 (Signal) darstellen. Diese Werte lassen sich für jedes Ereignis zu einer sogenannten Konfidenz mitteln. Durch einen Konfidenzschnitt lässt sich jedes Ereignis einer Klasse zuordnen. Dabei werden alle Ereignisse mit einer Konfidenz kleiner dem Schnittwert als Untergrund und alle anderen als Signal eingeordnet. Die so eingeordneten Ereignisse werden auch Untergrund- und Signalkandidaten genannt. Näheres zu Random Forests ist in [Bre01] zu finden.

2 Grundlagen

Zur Einschätzung der Qualität der Separation lassen sich die so genannte Reinheit R und die Effizienz P verwenden. R ist definiert als der Anteil der richtig als Signal klassifizierten Ereignisse an allen Signalkandidaten. P ist definiert als der Anteil der richtig als Signal klassifizierten Ereignisse an der Gesamtzahl von Signalereignissen.

Beim Random Forest kann es zur Überanpassung an die für das Training verwendeten Ereignissen kommen. Diese sorgt dafür, dass der Lerner noch nicht gesehene Ereignisse des gleichen Typs falsch zuordnet. Um eine Abschätzung für diese Schwankung der Separationsergebnisse des Random Forests auf unterschiedlichen Ereignismengen zu erhalten kann eine n -fache Kreuzvalidierung (KV) durchgeführt werden. Bei der KV wird die verwendete Ereignismenge in n gleich große Teile aufgeteilt. In jedem Schritt der KV werden $n - 1$ Teilmengen für das Trainieren eines Modells verwendet, das auf den übrig bleibenden Ereignissen validiert wird. Dieser Schritt wird n mal durchgeführt, so dass jede Teilmenge ein Mal zur Validierung verwendet wird. Durch Optimierung der Separationsqualität auf Ereignissen auf denen das Modell nicht trainiert wurde kann eine Überanpassung an statistische Schwankungen vermieden werden. [Koh+95]

3 Analyse

Ziel dieser Analyse ist ein Vergleich des DC- und EXT-Vetos anhand der Ereignisse, die diese Vtos passieren. Betrachtet werden die Differenzen der Raten an gefundenen Neutrinokandidaten und Myon-Neutrinokandidaten nach den Separationen, als auch die Korrelation ihrer rekonstruierten und wahren Energien.

Dafür werden drei Klassen von Ereignissen verwendet. Elektron- und Myon-Neutrino Datensätze auf Basis der IceCube Detektorkonfiguration aus dem Jahr 2013, die mit **GENIE** [And+10] erstellt wurden. Tau-Neutrinos werden in der Analyse nicht beachtet, da ihre Ereignissignatur der von Elektron-Neutrinos gleicht und für sie kein Modell für den Fluss vorhanden ist. Die Energie der simulierten Neutrinos liegt im Bereich von $1\text{ GeV} - 1000\text{ GeV}$. Sie werden nach einem Spektrum mit spektralem Index $\gamma = 2.0$ erzeugt und auf eines mit $\gamma = 2.7$ umgewichtet, was dem erwarteten Neutrinofluss entspricht. Dadurch erhält jedes Ereignis ein Gewichtsattribut, das nur zur Berechnung der finalen Qualitätsparameter und Raten verwendet wird. Zusätzlich werden **CORSIKA** [Hec+98] Myonen-Datensätze verwendet, die mit einem Hörandel Modell [Hör04] im Bereich einer Primärteilchenenergie von $600\text{ GeV} - 10^{11}\text{ GeV}$ produziert wurden.

Damit die Daten mit Methoden des maschinellen Lernens verarbeitet werden können, müssen sie erst vorprozessiert werden. Dabei werden Attribute entfernt, die konstant oder nicht in jedem Datensatz vorhanden sind, keine Informationen über die Physik der Ereignisse enthalten, oder einen Anteil an not a number (NaN) Werten von über 80% haben. NaNs treten unter anderem bei fehlgeschlagene Rekonstruktionen auf. Die NaN-Verteilung der Attribute ist im Anhang A.2 dargestellt. Zuletzt werden Attribute entfernt deren Pearson-Korrelationskoeffizienten mit mindestens einem der noch vorhanden Attribute größer als 0.95 ist. Dadurch ist eine höhere Stabilität der Analyse zu erwarten [RMS12]. Es werden die ungewichteten Pearson-Korrelationskoeffizienten verwendet, da sie für die verwendeten Daten weitestgehend mit den gewichteten Korrelationskoeffizienten übereinstimmen (siehe Anhang A.3). Von den anfänglich vorhandenen 2000 Attributen bleiben nach diesen Schritten 70 übrig.

Die Ereignisselektion ist in zwei Stufen aufgeteilt. Erst werden alle Neutrinos von Myonen getrennt. Dafür wird eine Attributselektion mittels mRMRe durchgeführt und dann mit einem Random Forest separiert. Für den Random Forest wird ein

3 Analyse

Konfidenzschnitt gewählt, der einen Kompromiss zwischen Reinheit des Signals und der Rate an Signalereignissen darstellt. Im zweiten Schritt werden Myon-Neutrinos von Elektron-Neutrinos getrennt, das Vorgehen ist wie im ersten Schritt. Diese zweischrittige Separationskette wird für jede Vetodefinition ein Mal durchgeführt. Da sich die in mRMRe für die Relevanz verwendeten Korrelationskoeffizienten durch Gewichtung nicht Signifikant verändern, wird hier auf die Gewichte der Ereignisse verzichtet. Der Random Forest wird ebenfalls ungewichtet trainiert, er erhält durch Gewichtung keine höhere Trennkraft. In der anschließenden Betrachtung der Qualitätsparameter und Konfidenzverteilungen werden die Gewichte beim Erstellen der Grafiken und Berechnen der Raten beachtet. So wird die Übertragbarkeit auf experimentelle Daten sichergestellt. Schließlich werden die Separationsqualitäten und Eigenschaften der Signalereignisverteilungen der beiden Vtos verglichen.

In dieser Analyse wird von Signal- und Untergrundereignissen, die nach einer Separation vorhanden sind, geschrieben. Diese Einteilung ist nur auf Simulationsdaten möglich, auf experimentellen Daten sind nur *Ereigniskandidaten* für Signal und Untergrund bekannt.

3.1 Neutrinoselection

Im ersten Separationsschritt wird das Signal in Form von beiden in den Datensätzen vorhandenen Neutrinoflavours vom Myon-Untergrund getrennt. Unterschieden wird dabei im wesentlichen die unterschiedliche Topologie der Ereignistypen. Das Signal besteht aus einer gleichen Anzahl von simulierten Elektron- wie Myon-Neutrinos. Diese Vorseparation sorgt für eine bessere Trennung als eine Separation mit drei Klassen oder eine direkte Separation der Myon-Neutrinos vom gesamten Untergrund.

3.1.1 Selektion

Die Attributselektion wird, wie in Abschnitt 2.3 beschrieben, mittels mRMRe in R [R D08] durchgeführt. Dabei wird eine gleiche Anzahl von simulierten Untergrund- wie Signalereignissen verwendet, was bedingt durch die Menge an vorhandenen CORSIKA Daten zu einer Beschränkung auf insgesamt 52 000 Ereignisse führt. Um die Stabilität der Selektion zu bestimmen werden durch Ziehen mit Zurücklegen fünf Datensätze gezogen, auf denen je ein Mal selektiert wird. Die fünf Stabilitätsverteilungen sind gemittelt in 3.1 dargestellt.

Es sind keine großen Ausreißer zu sehen und mit zunehmender Zahl an selektierten Attributen k scheint die Stabilität mit 0.95 in eine Sättigung über zu gehen. Die

Standardabweichung der Jaccard-Indices sinkt mit größer werdendem k und bleibt ab $k = 42$ etwa konstant. Es ist zu beachten, dass die Punkte für $k = 1, 2, 3$ eine Standardabweichung vom Mittelwert von Null haben. Bei $k = 1$ wird für jede Selektion ein anderes Attribut ausgewählt, es gibt also kein bestes Attribut. Für die folgende Separation werden 44 Attribute verwendet. Bei dieser Attributanzahl ist die Standardabweichung der Stabilität fast minimal und ihr Mittelwert nahe am Maximum. Für die endgültige Attributwahl wird aus den Attributmengen aus dem Stabilitätstest ein einziger Satz an Attributen generiert. Dafür werden für jedes Attribut dessen Ränge aus den Selektionen addiert und nach diesen summierten Rangwerten sortiert. Es werden nur die k (hier 44) höchsten Attribute weiter verwendet.

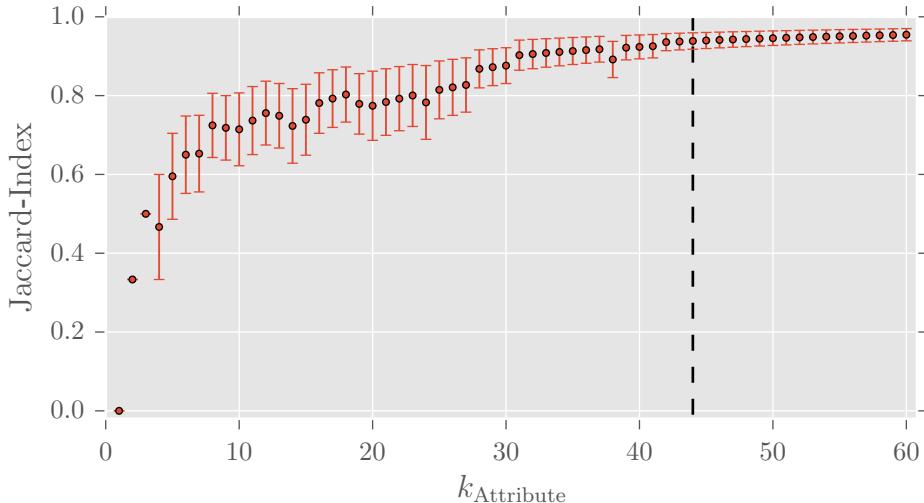


Abbildung 3.1: Aufgetragen sind die gemittelten Jaccard-Indices mit ihren Standardabweichungen gegen die Anzahl an selektierten Attributen.

3.1.2 Separation

In diesem Schritt werden die Neutrinos beider vorhandener Flavour von Myonen separiert. Es wird eine Separation für jede Vetoregion durchgeführt, wobei nur Ereignisse verwendet werden, die das jeweilige Veto passiert haben. Beide Separationen laufen gleich ab, vollständig beschrieben wird nur die des DC-Vetos. Die Konfidenzverteilungen und Qualitätsparameter der EXT-Separation sind im Anhang A.5 zu finden, da sie analog durchgeführt werden. Verwendet wird der WEKA Random Forest [Hal+09] in RapidMiner [Mie+06]. Um die Rechenzeit zu begrenzen wird jeder Forest aus 50 Bäumen aufgebaut. Für die Größe N der Attributmenge jedes

3 Analyse

Baums wird $N = \text{int}(\log(k) + 1)$ gewählt [Bre01]. Mit der Gesamtzahl an Attributen k aus der mRMRe Selektion. Es werden alle 26 000 vorhandenen CORSIKA Untergrundereignisse verwendet und je 250 000 Ereignisse pro verwendetem Neutrinoflavour. Für Neutrinozahlen unter 100 000 ordnet der Random Forest jedes Ereignis als Untergrund ein. Es ist anzunehmen, dass die Separationsqualität durch die geringe Menge an Untergrundereignissen begrenzt wird. Um eine Abschätzung der Separationsqualität und ihrer Schwankung auf ungesesehenen Daten zu erhalten, wird eine fünffache Kreuzvalidierung durchgeführt. Da dabei nicht alle Ereignisse Verwendung finden und das Signal zu Untergrund Verhältnis beim Trainieren nicht dem realen entspricht, werden die jeweiligen Raten durch den jeweils verwendeten Anteil geteilt, um die real zu erwartenden Raten zu erhalten.

In Abbildung 3.2 sind die Ereignisraten für die Konfidenz, dass ein Ereignis zum Signal gehört, dargestellt. Sie zeigt ein starkes Überlappen der beiden Verteilungen und damit eine niedrige Trennkraft. Für Myonen hat die Rate ihr Maximum bei niedrigen Konfidenzwerten um 0.1. Die Neutrinos haben ihr Maximum im selben Bereich. Von dort bis zur Konfidenz 0.8 sinken sie auf ein Viertel ihres Maximums, danach fallen sie um mehrere Größenordnungen ab. Ein Großteil der Neutrinos ist also für den Lerner nicht von Myonen zu unterscheiden. Die Ausreißer der Myonen mit Konfidenzen größer 0.6 entsprechen jeweils einem klassifizierten Ereignis.

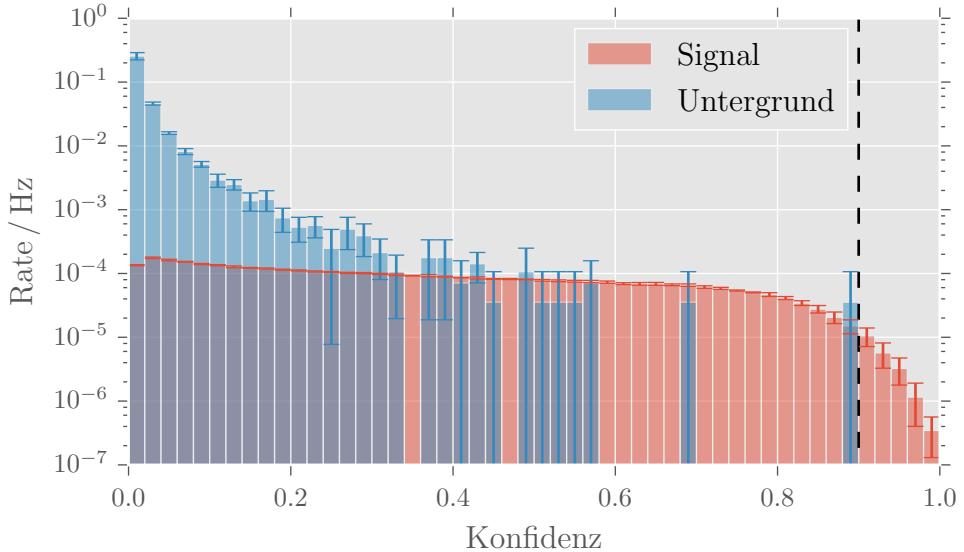


Abbildung 3.2: Histogramme der Konfidenzverteilungen der gemittelten Ereignisraten von Myonen (Untergrund) und Neutrinos (Signal) in Hertz. Die Standardabweichungen aus der fünffachen Kreuzvalidierung sind als Fehlerbalken dargestellt. Der verwendete Konfidenzschnitt ist als vertikale Linie bei 0.9 eingezeichnet.

In Abbildung 3.3 sind die gemittelten Reinheiten und Effizienzen aus der Kreuzvalidierung dargestellt. Bei einer Konfidenz von 1.0 ist für die Reinheit kein Wert vorhanden, es kommt hier zu einem Rundungsfehler und damit zu einem NaN-Wert. Die Reinheiten zeigen für den Übergangsbereich zwischen den Konfidenzen 0.3 und 0.9 hohe Standardabweichungen von ihren Mittelwerten. Grund für die vermeintlich perfekte Reinheit von 1.0 für Konfidenzen größer 0.9 ist die geringe Statistik des Myondatensatzes. Bei einer größeren Anzahl an Myonereignissen ist zu erwarten, dass die Konfidenzverteilung der Myonen stetiger wird und für höhere Konfidenzwerte mehr Myonen auftauchen. Über einer Konfidenz von 0.9 werden die Effizienzen kleiner als 1%. Als Konfidenzschnitt für die weitere Analyse wird der Wert 0.9 genommen. Dies ist der niedrigste Wert für den keine Untergrundereignisse mehr vorhanden sind, womit die Reinheit maximiert wird und als Nebenbedingung die Effizienz maximiert wird. Beim EXT-Veto sind Myonereignisse nur bis zu einer Konfidenz von 0.84 vorhanden. Um den Vergleich einheitlicher zu machen wird dort der selbe Konfidenzschnitt bei 0.9 angesetzt.

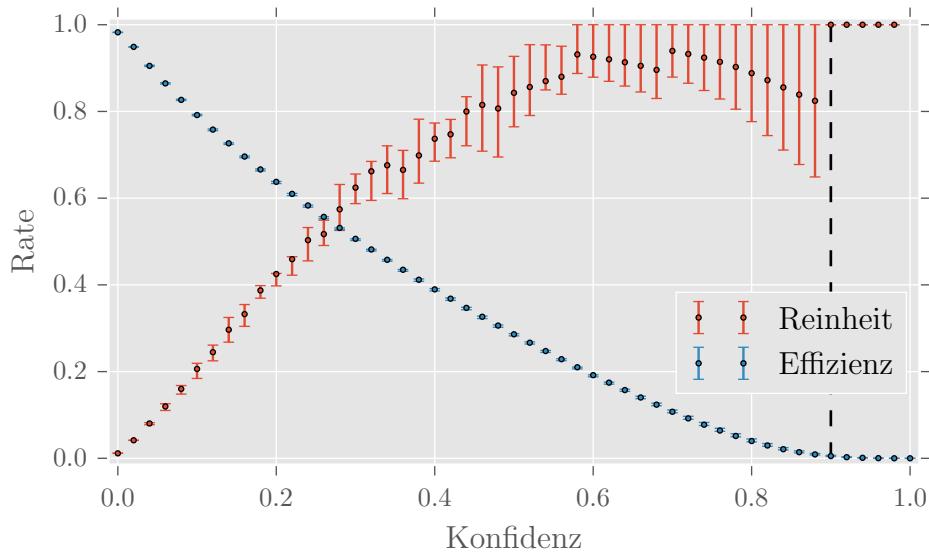


Abbildung 3.3: Aufgetragen sind die gemittelten Qualitätsparameter Reinheit (rot) und Effizienz (blau) mit ihren Standardabweichungen gegen die Konfidenz.

3.2 Myon-Neutrinoselection

In dieser Selektion sind Myon-Neutrinos das Signal und Elektron-Neutrinos der Untergrund. Von beiden Teilchensorten werden nur solche Ereignisse verwendet, die nach der Neutrino-separation übrig geblieben sind.

3.2.1 Selektion

Das Vorgehen ist gleich dem der Selektion in 3.1.1, es wird nur eine größere Zahl von 500 000 Ereignissen verwendet.

Die Stabilitätsverteilung der Attributselektion ist in 3.4 dargestellt. Sie zeigt im Vergleich zur vorherigen Stabilitätsverteilung größere Schwankungen. Die Standardabweichungen der gemittelten Stabilitäten sind im Vergleich zur vorherigen Selektion größer. Ab $k \sim 23$ werden die Standardabweichungen im Vergleich zu niedrigeren k -Werten kleiner und finden bis auf einige Ausreißer in den höchsten k -Werten ihr Minimum um $k = 36$. Bei $k = 1$ ist die Stabilität konstant Eins, es gibt also ein bestes Attribut. Nach $k = 1$ findet ein Abfallen statt, worauf bei $k = 6$ ein Anstieg folgt, der erste Sättigung bei $k = 23$ zeigt, und ein lokales Maximum um $k = 36$. Bei $k > 41$ findet wieder eine Sättigung statt, die nahe an eine Stabilität von 1 kommt. Für die anschließende Separation werden 36 Attribute verwendet. Hier haben Stabilität und ihre Standardabweichung ein lokales Maximum bzw. Minimum, das gleich oder nahe dem globalen ist.

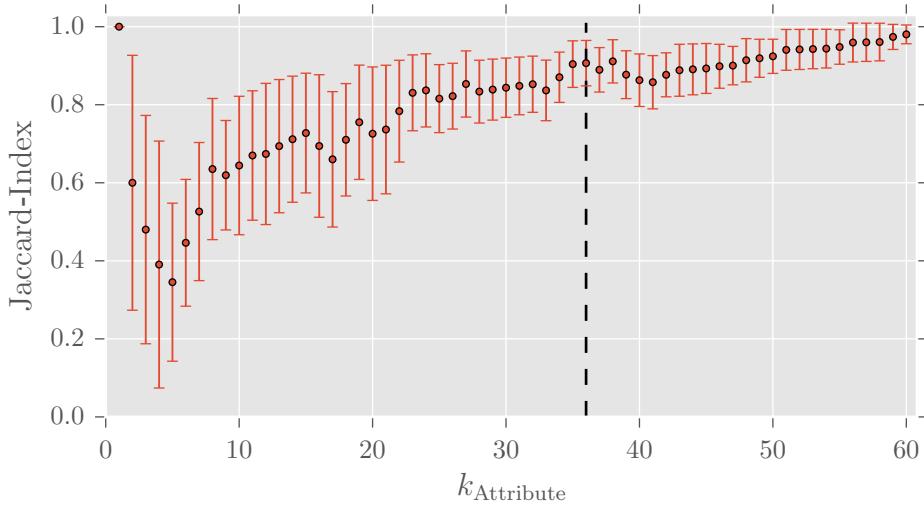


Abbildung 3.4: Aufgetragen sind die gemittelten Jaccard-Indices mit ihren Standardabweichungen gegen die Anzahl an selektierten Attributen.

3.2.2 Separation

Die Separation läuft analog zur Neutrinoseparation. Nachdem alle Myonereignisse durch die Neutrinoseparation entfernt wurden, werden jetzt Myon-Neutrinos (Signal) von Elektron-Neutrinos (Untergrund) getrennt. Für das Trainieren des Lerners stehen beim DC-Veto 54 077 und beim EXT-Veto 10 716 Ereignisse zur Verfügung, was die erreichbare Separationsqualität einschränkt.

Die gemittelten Konfidenzverteilungen mit ihren Standardabweichungen aus der fünfachen Kreuzvalidierung sind in Abbildung 3.5 dargestellt. Die Signal- und Untergrund-Verteilungen überlappen sich stark, mit Maxima um einer Konfidenz von 0.7. Für Konfidenzen kleiner 0.4 ist die Untergrundrate höher als die Signallrate, wobei die Abweichungen teilweise innerhalb einer Standardabweichung der Mittelwerte liegt. Für Konfidenzen größer 0.6 ist die Signallrate um mehrere ihrer Standardabweichungen höher als die Untergrundrate.

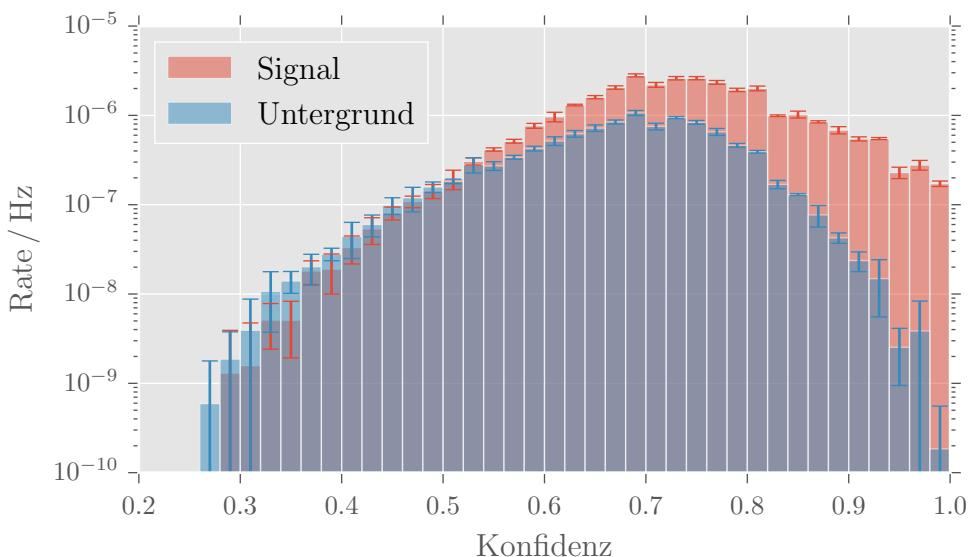


Abbildung 3.5: Histogramme der Konfidenzverteilungen der gemittelten Ereignisraten von Elektron-Neutrinos (Untergrund) und Myon-Neutrinos (Signal) in Hertz. Die Standardabweichungen aus der fünfachen Kreuzvalidierung sind als Fehlerbalken dargestellt. Ereignisse mit Konfidenzen unter 0.2 sind nicht vorhanden.

In Abbildung 3.6 sind die gemittelten Reinheiten und Effizienzen mit ihren Standardabweichungen aus der Kreuzvalidierung dargestellt. Für Konfidenzen unter 0.5 ist bei der Reinheit das nach der Neutrinoseparation vorhandene Signal zu Untergrund Verhältnis von etwa 7:4 zu sehen, worüber hinaus noch keine Trennung

3 Analyse

erfolgt. Ab Konfidenzen von 0.6 steigt die Reinheit an, verbunden mit einem steilen Abfallen der Effizienz. Soll eine Reinheit von 99% erreicht werden müssen 98% der Signalereignisse verworfen werden. Der Konfidenzschnitt wird bei 0.84 ange setzt. Hier hat das DC-Veto eine Reinheit von $(95.3 \pm 0.8) \%$ und eine Effizienz von $(10.9 \pm 0.3) \%$

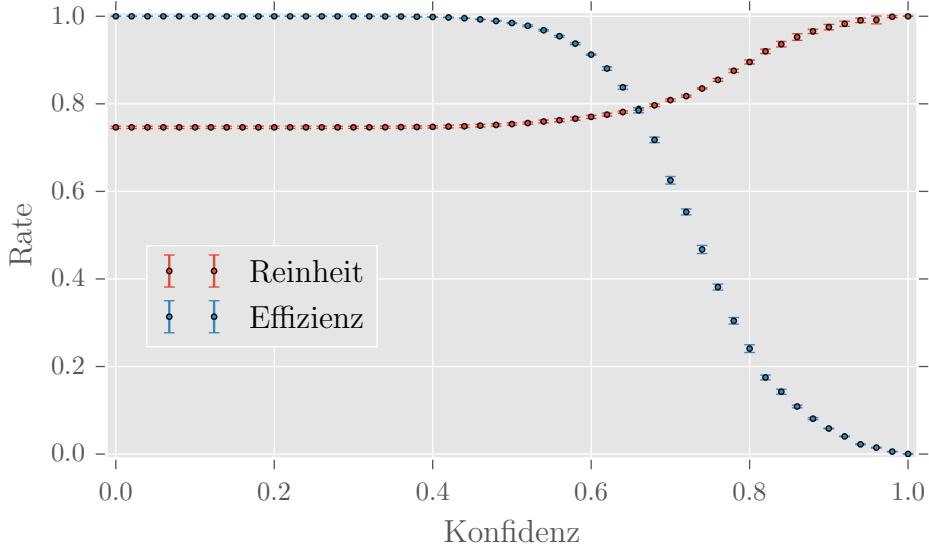


Abbildung 3.6: Aufgetragen sind die gemittelten Qualitätsparameter Reinheit (rot) und Effizienz (blau) mit ihren Standardabweichungen gegen die Konfidenz.

3.3 Vergleich der Vetoregionen

Verglichen werden die Separationen aus den Abschnitten 3.1.2 und 3.2.2. Betrachtet werden die Ereignisraten im Vergleich zur Konfidenz und der Zusammenhang der rekonstruierten Energie mit der wahren Energie der Ereignisse. Als rekonstruierte Energie wird der Wert des Attributs `MPEFitMuEX` verwendet. Es berechnet die Energie der Ereignisse über den Energieverlust dE/dx der Myonen. Es ist in allen Separationen als Attribut vorhanden und hat nach dem Konfidenzschnitt von 0.9 in der Neutrinoselection einen NaN-Anteil von 2%. Diese NaN-Ereignisse werden im Vergleich der Energien nicht verwendet. Die im Folgenden betrachteten relativen Ratendifferenzen Δ sind definiert über:

$$\Delta := \frac{r_{\text{DC}} - r_{\text{EXT}}}{r_{\text{DC}}} \quad (3.1)$$

Mit den Ereignisraten r für das jeweilige Veto.

3.3.1 Neutrinoselection

Abbildung 3.7 zeigt die Δ von DC- und EXT-Veto für Werte der wahren Energie und Konfidenz aus der Neutrinoselection. Es ist zu sehen, dass die Raten des EXT-Vetos für Konfidenzen unter 0.4 größer als die des DC-Vetos sind. Die dort höheren Raten sind jedoch ohne Nutzen, da in der Myon-Neutrinoseparation nur Ereignisse mit einer Konfidenz größer 0.9 verwendet werden. Bei Konfidenzen größer 0.6 verfügt das DC-Veto über höhere Ereignisraten, während sich zu 0.5 hin von beiden Seiten die Raten angleichen. Diese Eigenschaften der Verteilung treten in jedem Energiebereich auf, wobei der Betrag der Δ mit steigender Energie zunimmt. Für Energien größer 500 GeV und Konfidenzen größer 0.9 kommt es zu ungefüllten Bins; es sind zu wenig Ereignisse vorhanden um den gesamten Parameterraum abzudecken.

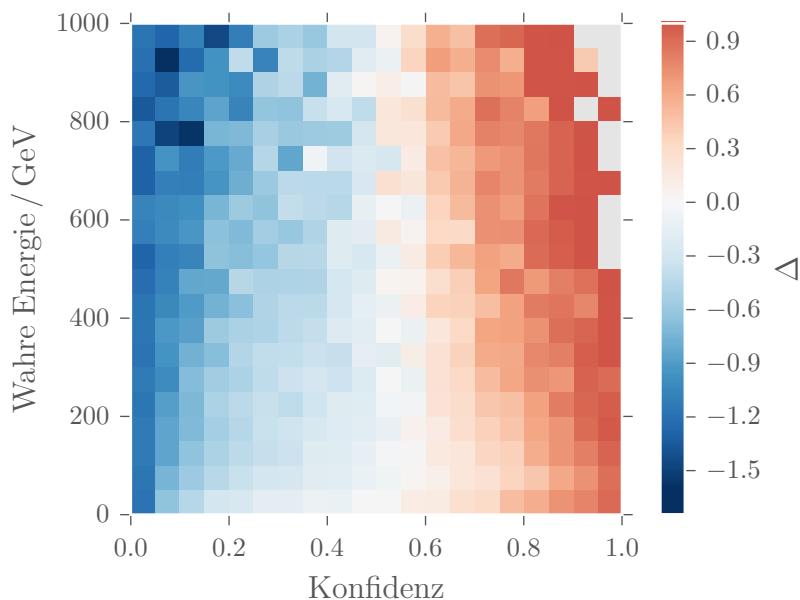


Abbildung 3.7: Aufgetragen ist wahre Energie gegen Konfidenz für alle Neutrinoereignisse. Die relativen Differenzen der Ereignisraten von DC- und EXT-Vetos sind farblich kodiert.

In Abbildung 3.8 sind die Δ gegen rekonstruierte und wahre Energie nach dem Konfidenzschnitt bei 0.9 zu sehen. Wahre Energien größer 200 GeV und rekonstruierte Energien größer 800 GeV werden nicht betrachtet, da dort insgesamt maximale Raten von nur etwa 0.02 μ Hz vorhanden sind. Von vier Bins abgesehen, zeigt das DC-Veto in jedem Energiebereich höhere Raten.

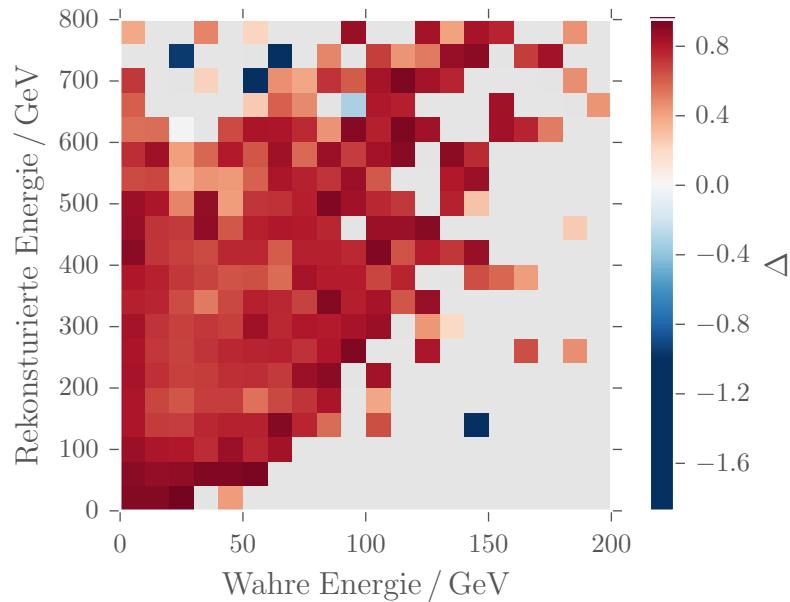


Abbildung 3.8: Aufgetragen ist wahre Energie gegen rekonstruierte Energie nach dem Konfidenzschnitt bei 0.9. Die Δ von DC- und EXT-Vetos sind farblich kodiert.

Abbildung 3.9 zeigt die in Abbildung 3.8 zu Grunde liegenden absoluten Raten für das DC-Veto. Um die Korrelation von rekonstruierter und wahrer Energie über alle Energiebreiche sichtbarer zu machen wurden die Raten um den spektralen Index $\gamma = 2.7$ korrigiert. Dazu wird jedes Ereignis auf $w_{\text{neu}} = w_{\text{alt}} E_{\text{wahr}}^\gamma$ neu gewichtet. Die schwarz gestrichelte Linie mit Steigung Eins zeigt den für eine perfekte Energierekonstruktion zu erwartende Häufungszone der Raten. Unter ihr befindet sich nur ein Anteil von weniger als 1% der Gesamtrate. Die rekonstruierten Energien werden also kaum unterschätzt. Die Häufungszone der Raten ist stark über die Linie verschoben, die Energien werden im Durchschnitt um das Achtfache überschätzt. Die mit w_{alt} gewichtete Pearson-Korrelation zwischen den rekonstruierten und wahren Energien ergibt sich zu $\rho_{\text{DC}} = 0.70$. Wird die Korrelation mit w_{neu} gewichtet, ergibt sich $\rho_{\text{DC},\text{neu}} = 0.76$. Die Korrektur verstärkt die Korrelation anstatt nur die Energiebereiche vergleichbarer zu machen. Beim EXT-Veto (Abbildung A.9) besteht mit $\rho_{\text{EXT}} = 0.67$ eine niedrigere Korrelation.

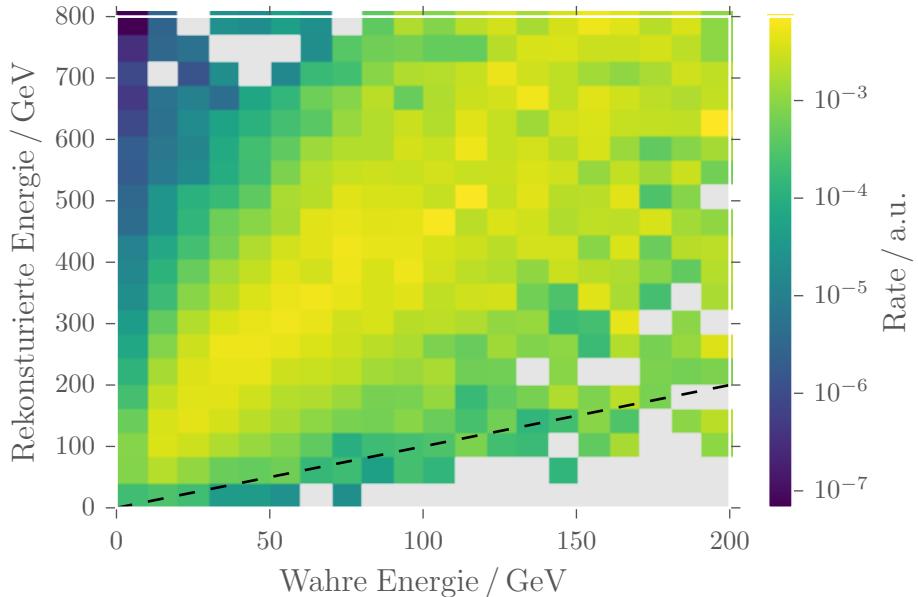


Abbildung 3.9: Aufgetragen ist wahre Energie gegen rekonstruierte Energie nach dem Konfidenzschnitt bei 0.9 für die Neutrinoselection mit dem DC-Veto. Die absoluten Ereignisraten sind farblich kodiert.

3.3.2 Myon-Neutrinoselection

In Abbildung 3.10 sind die Δ von DC- und EXT-Veto für Werte der wahren Energie und Konfidenz dargestellt. Verwendet werden nur Signalereigniskandidaten nach Anwenden der Random Forests aus der Myon-Neutrinoselection.

Für die Konfidenzen werden die Klassifikationen der Random Forests aus der Kreuzvalidierung der Myon-Neutrinoselection verwendet. Jeder Random Forest klassifiziert dabei nur Ereignisse, die er noch nicht gesehen hat. Anwenden eines einzigen Forests auf alle Ereignisse ist ohne Einführung eines Bias in den Ergebnissen nicht möglich, da jeder Forest auf $\frac{4}{5}$ der vorhandenen Daten trainiert wurde. Die Raten des DC-Vetos dominieren bis auf vereinzelte Ausnahmen in jedem Energie- und Konfidenzbereich. Gleichermaßen zeigt sich bei den Raten gegen rekonstruierte und wahre Energie in Abbildung 3.11. Hier werden alle Signalereigniskandidaten nach dem Konfidenzschnitt bei 0.84 verwendet. Die Korrelationen zwischen rekonstruierten und wahren Energien sind $\rho_{DC} = 0.67$ und $\rho_{EXT} = 0.63$. Sie sind im Vergleich zur Neutrinoselection gesunken.

Tabelle 3.1 fasst die Myon-Neutrino-Raten nach den Schritten der Analyse noch einmal zusammen. Zum Nachteil des EXT-Vetos wurde in der Neutrinoseparation

3 Analyse

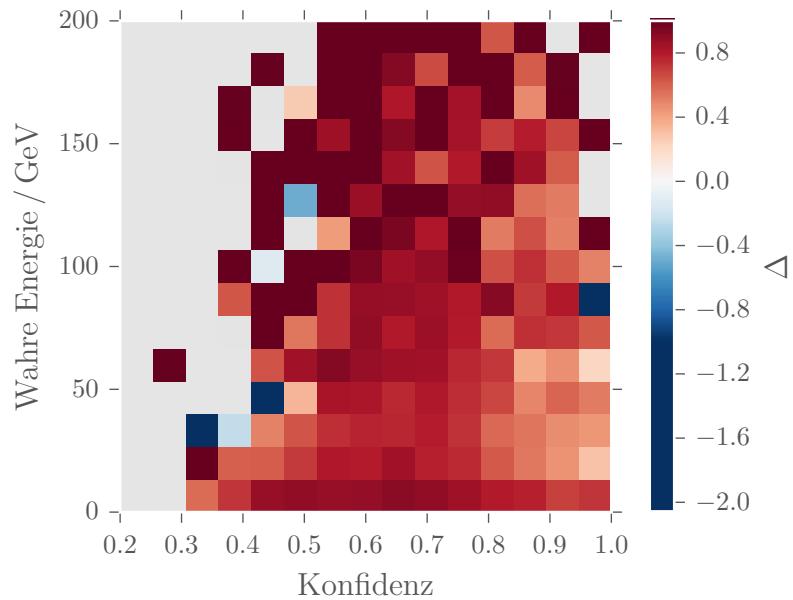


Abbildung 3.10: Aufgetragen ist wahre Energie gegen Konfidenz. Die Δ von DC- und EXT-Vetos sind farblich kodiert.

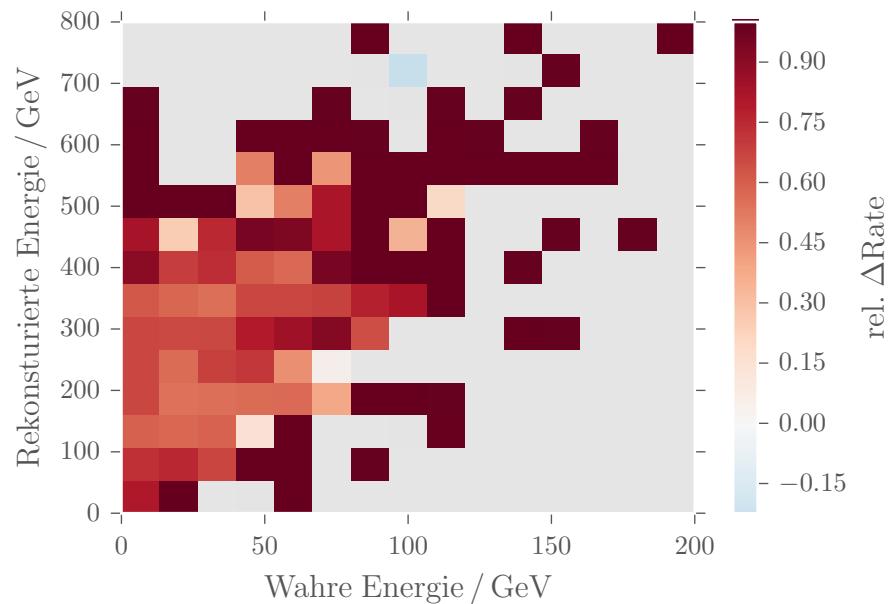


Abbildung 3.11: Aufgetragen ist wahre Energie gegen rekonstruierte Energie. Die Δ von DC- und EXT-Veto sind farblich kodiert.

3.3 Vergleich der Vetoregionen

der Konfidenzschnitt für beide Vetos bei 0.9 angesetzt, obwohl beim EXT-Veto Untergrundereignisse nur bis zu einer Konfidenz von 0.84 zu finden sind. Wird für das EXT-Veto der Schnitt bei 0.84 angesetzt, bleibt eine Neutrinoereignisrate von $28.43 \mu\text{Hz}$ übrig. Die höhere Rate vor der ersten Separation beim EXT-Veto sorgt in diesem Fall für keine höheren Raten nach der Separation.

Tabelle 3.1: Die Ereignisraten und -zahlen an Myon-Neutrinos nach den einzelnen Separationsschritten für DC- und EXT-Veto.

	Ereigniszahlen		Ereignisraten / μHz	
	DC	EXT	DC	EXT
Veto Filter	3 401 258	4 389 350	2995.6	3889.9
Neutrinoelektion	35 279	7188	30.4	6.2
Myon-Neutrinoelektion	6013	2191	4.1	1.5

4 Zusammenfassung und Ausblick

Zusammenfassung

In dieser Arbeit werden zwei gleich aufgebaute zweischrittige Separationen zur Selektion von atmosphärischen Myon-Neutrino-Kandidaten durchgeführt und ihre Ergebnisse verglichen. Die dafür notwendigen Ereignisse entstehen jeweils durch Filtern mit unterschiedlichen Vetoregionen, dem DC- und dem EXT-Veto, die alle Ereignisse verwerfen, deren Interaktion vermutlich im vom Veto definierten Vetovolumen stattfindet. Verwendet werden dafür Simulationsdaten von Myon-, Elektron-Neutrino und Myon-Neutrino Ereignissen mit Attributen wie Energie- und Richtungsrekonstruktionen. Zuerst wird die jeweilige Vetobedingung auf alle vorhandenen Ereignisse angewendet. Dann werden mittels mRMRe 44 Attribute für die Separation von Myon- und Neutrinoereignissen ausgewählt. In der Neutrinoseparation wird in einer fünffachen Kreuzvalidierung ein Random Forest mit 50 Bäumen auf allen vorhandenen Myonereignissen und jeweils 250 000 Elektron- sowie Myon-Neutrino-Ereignissen trainiert. Dieser wird auf alle Simulationsdaten angewendet. Alle Ereignisse mit einer Konfidenz unter 0.9 werden dem Untergrund zugeordnet und verworfen, alle anderen sind Signalereigniskandidaten. Für das DC-Veto bleiben nach diesem Schritt 54 077 simulierte Ereignisse mit einer Gesamtrate von $40.76 \mu\text{Hz}$ übrig, für das EXT-Veto bleiben 10 716 Ereignisse mit einer Rate von $8.01 \mu\text{Hz}$ übrig. Unter den selektierten Ereignissen sind nur noch Elektron- und Myon-Neutrinos. Sie werden in einer zweiten Separation voneinander getrennt. Dafür werden mittels mRMRe 36 Attribute selektiert und in einer fünffachen Kreuzvalidierung je ein Random Forest mit 200 Bäumen trainiert. Die Ereignisse werden so in Myon-Neutrino- (Signal) und Elektron-Neutrino-Kandidaten (Untergrund) separiert. Der Konfidenzschnitt wird bei 0.84 gesetzt, was beim DC-Veto zu einer Signalreinheit von $(95.3 \pm 0.8)\%$ führt. Für das DC-Veto bleiben Signalereigniskandidaten mit einer Gesamtrate von $4.63 \mu\text{Hz}$ übrig, für das EXT-Veto $1.67 \mu\text{Hz}$. Auf Basis der Raten an verbleibenden Signalereigniskandidaten wird ein Vergleich der beiden Vetoregionen durchgeführt. Nach dem ersten Separationsschritt werden die Neutrinoereignisraten zusammen mit der Konfidenz und wahren Energie der Ereignisse verglichen. Es zeigt sich, dass das DC-Veto für Konfidenzen größer 0.64 in jedem Energiebereich über höhere Ereignisraten als das EXT-Veto verfügt. Beim Vergleich der rekonstruierten und wahren Energie der Ereignisse ergibt sich für die Vtos eine Korrelation von $\rho_{\text{DC}} = 0.7$

und $\rho_{\text{EXT}} = 0.67$, sowie ein Überschätzen der wahren Energie. Danach werden die Raten von Kandidaten für Myon-Neutrino-Ereignisse nach der zweiten Separation verglichen. Die Raten des DC-Vetos sind für jeden Konfidenz- und Energiebereich höher als die des EXT-Vetos.

Ausblick

Weiterführend kann die in dieser Arbeit aufgestellte Separations- und Analysekette automatisiert werden, um weitere Vetodefinitionen und ihre Effekte auf die Qualität der Separation zu testen. Zudem sollte ein Vergleich der Simulationsdaten mit experimentellen Daten durchgeführt werden und Attribute mit zu geringer Übereinstimmung entfernt werden. Damit wird sichergestellt, dass die Separation auf experimentellen Daten ähnliche Ergebnisse liefert, wie in der Kreuzvalidierung. Zusätzlich kann die durchgeführte Separation weiter verbessert werden. Um die Trennkraft zu erhöhen können neue, auf den Niederenergiebereich angepasste, Attribute erstellt werden. Anstatt die Vetoentscheidungen wie bisher als Filter für Ereignisse zu verwenden, können sie dem Lerner als Attribut übergeben werden.

Bei der Neutrinoelektion gehen die meisten Signalereignisse verloren. Hauptursache dafür ist wahrscheinlich die geringe Anzahl an **Corsika** Untergrundereignissen. Um dem entgegen zu wirken können mehr Myonereignisse simuliert und zum Trainieren verwendet werden. Es können auch experimentelle Daten als Myonuntergrund verwendet werden. Möglich ist dies durch das hohe Signal-Untergrund-Verhältnis von $1:10^3$. Der geringe Anteil an Neutrinos in solch einem Untergrunddatensatz beeinträchtigt die Separationsqualität des Random Forests durch seine Robustheit gegenüber falsch gelabelten Ereignissen nur in geringem Maße. Um Stabilität und Trennkraft der Separation zu erhöhen kann eine höhere Zahl von Bäumen trainiert werden. Auf Grund von Limitationen der Rechenzeit waren die in dieser Arbeit durchgeführten Neutrinoseparationen auf 50 Bäume begrenzt. Die Separation der Myon-Neutrinos sollte sich durch eine mehrfache Separation für unterschiedliche Energiebereiche verbessern lassen. Eine so optimierte Separationskette kann auf aktuelle experimentelle Daten angewendet werden, um etwa eine Bestimmung des Myon-Neutrinoflusses zu ermöglichen.

A Anhang

A.1 Verwendete Software

Die Vorprozessierung der im HDF5 [TheNN] Format vorliegenden Daten wurde mit verschiedenen Python [RB91] Paketen durchgeführt. H5py [Col+08] zum Einlesen und erneuten Schreiben der Daten. Numpy [Oli+] und Pandas [McK+] zum Anwenden von mathematischen Funktionen als auch zum Transformieren der Daten. Die Attributselektion wurde mit der R [R D08] Bibliothek mRMRe [De +13] in Python über den R zu Python wrapper rpy2 [Gau+09] durchgeführt. Die Separation wurde mit dem WEKA Random Forest [Hal+09] in RapidMiner [Mie+06] durchgeführt. Alle Grafiken wurden mit Matplotlib [Hun07] erstellt.

A.2 NaN-Verteilung

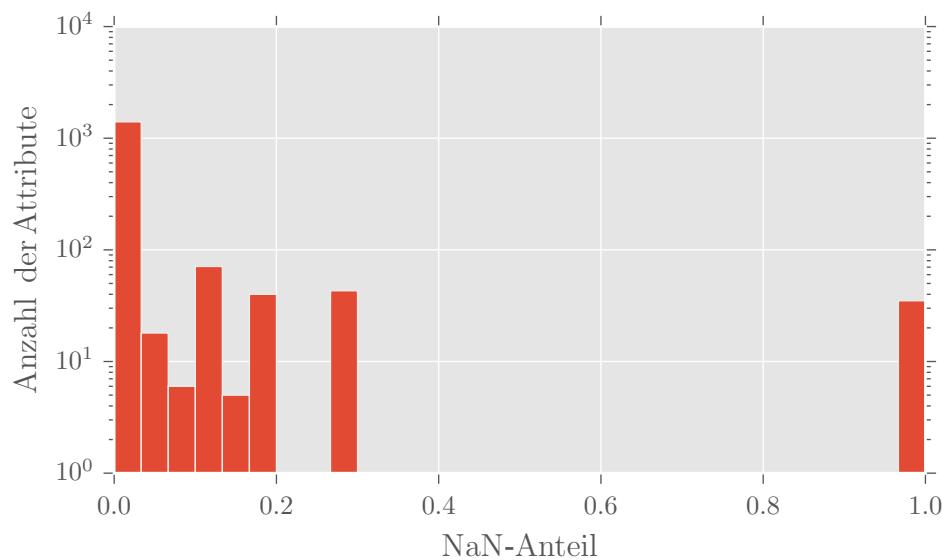


Abbildung A.1: Histogramm der NaN-Anteile.

A.3 Vergleich von gewichteten und ungewichteten Korrelationskoeffizienten

Die Ereignisse sind je nach ihrer Energie unterschiedlich gewichtet. Weiter kann nicht ausgeschlossen werden, dass die einzelnen Attribute für verschiedene Energiebereiche unterschiedlich stark miteinander oder mit dem Zielattribut korreliert sind. Diese sich verändernden Korrelationen würden für eine Verfälschung der Attributsauswahl sorgen. Daher wird untersucht, ob die Gewichtung der Ereignisse die Korrelationskoeffizienten beeinflusst. Zum Vergleich werden die gewichteten und ungewichteten Pearson-Korrelationskoeffizienten der Attribute verwendet.

In Abbildungen A.2 und A.3 ist zu sehen, dass die relativen Abweichungen der beiden Arten von Koeffizienten nur gering sind. Die gewichteten und ungewichteten Koeffizienten sind also weitestgehend äquivalent.

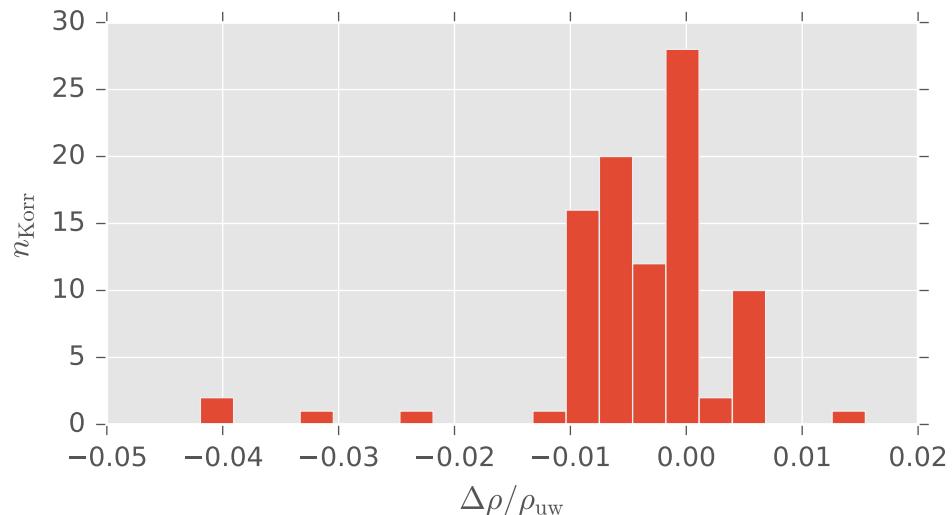


Abbildung A.2: Histogramm der relativen Differenzen zwischen gewichteten und ungewichteten Pearson-Korrelationskoeffizienten der Attribute mit der wahren Energie.

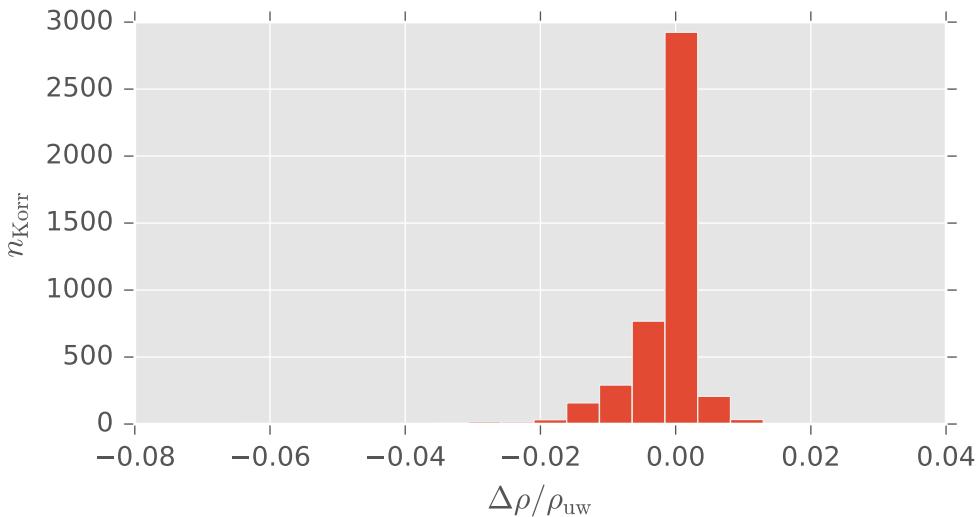


Abbildung A.3: Histogramm der relativen Differenzen zwischen gewichteten und ungewichteten Pearson-Korrelationskoeffizienten aller möglichen Attributskombinationen.

A.4 Vergleich von mRMR Methoden anhand von Testseparationen

Verglichen wird die in Abschnitten 3.1.1 und 3.2.1 verwendete mRMRe Variante und die mRMR der RapidMiner Feature Selection Extension [Sch10] mit dem MID Kriterium (weiter FSE-mRMR genannt). Dafür werden fünfach kreuzvalidierte Separationen von Elektron- und Myon-Neutrinos mit einem Random Forest durchgeführt. Bei beiden Separationsketten unterscheiden sich nur die Arten der Attributselektionen. Die gemittelten relativen Abweichungen der Reinheiten und Effizienzen sind in Abbildung A.4 dargestellt. Positive Werte stehen für eine höhere Qualität der Separation auf Basis der Attributsmenge der mRMRe Methode im Vergleich zur FSE-mRMR. Sie sind bis zu Konfidenzwerten von 0,4 mit Null verträglich. Danach weichen sie um bis zu 5% voneinander ab, die Standardabweichungen werden teilweise so groß wie die relativen Abweichungen selbst. Die relative Reinheit der mRMRe ist höher und ihre relative Effizienz niedriger. Die mRMRe Methode liefert also vergleichbare Ergebnisse wie die FSE-mRMR Varianten.

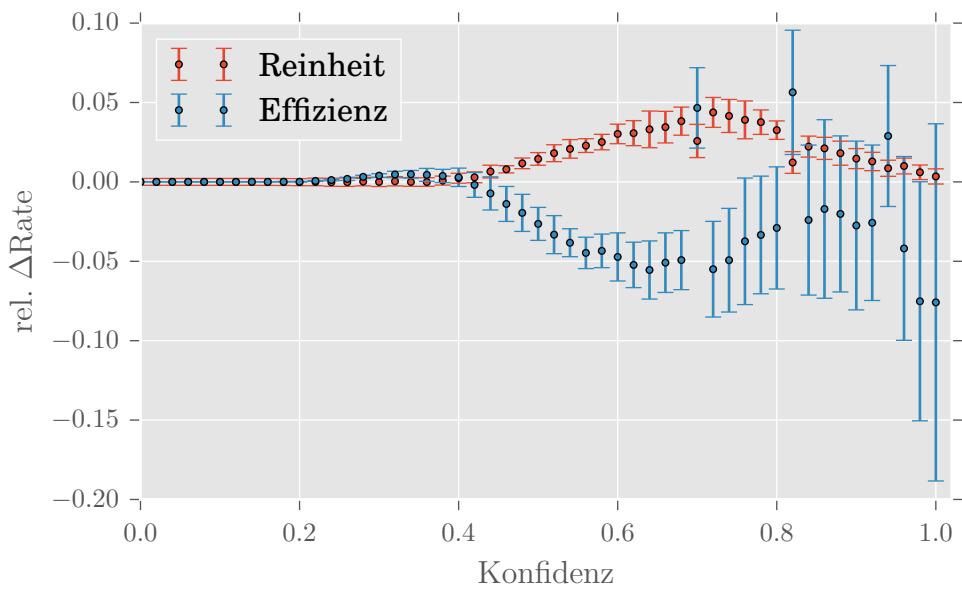


Abbildung A.4: Aufgetragen sind die gemittelten relativen Differenzen der Qualitätsparameter Effizienz und Reinheit zweier Separationen mit vorheriger Attributselektion durch mRMRe und FSE-mRMR. Positive Werte stehen für eine bessere Separation mit der mRMRe Attributmenge. Fehlerbalken sind die Standardabweichungen der Mittelwerte.

A.5 Separationsqualität der EXT-Veto Neutrinoselection

Dieses Kapitel beinhaltet der Vollständigkeit halber Grafiken für das EXT-Veto, die in Abschnitt 3.1.2 und 3.2.2 für das DC-Veto dargestellt wurden.

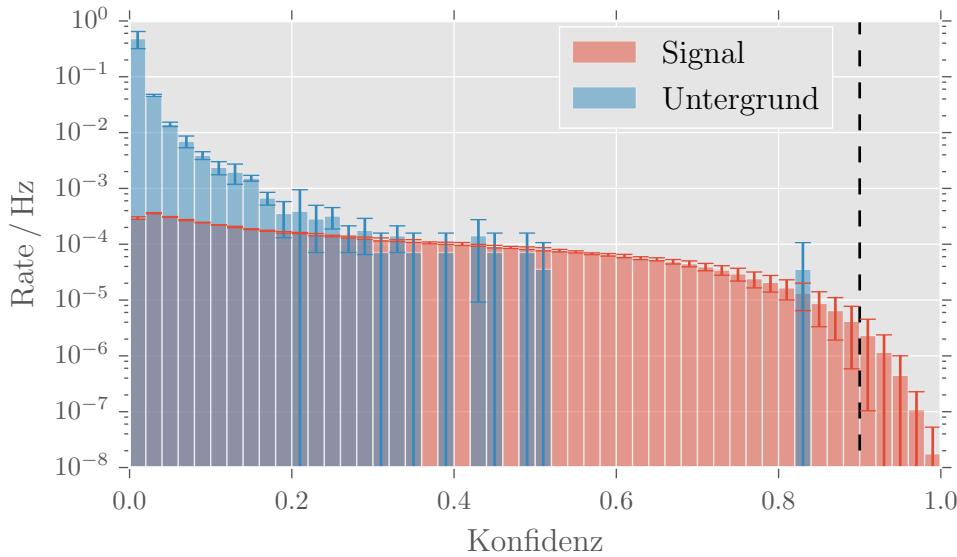


Abbildung A.5: Aus der Neutrino-separation mit dem EXT-Veto. Histogramme der Konfidenzverteilungen der gemittelten Ereignisraten von Myonen (Untergrund) und Neutrinos (Signal) in Hertz. Die Standardabweichungen aus der fünffachen Kreuzvalidierung sind als Fehlerbalken dargestellt.

A Anhang

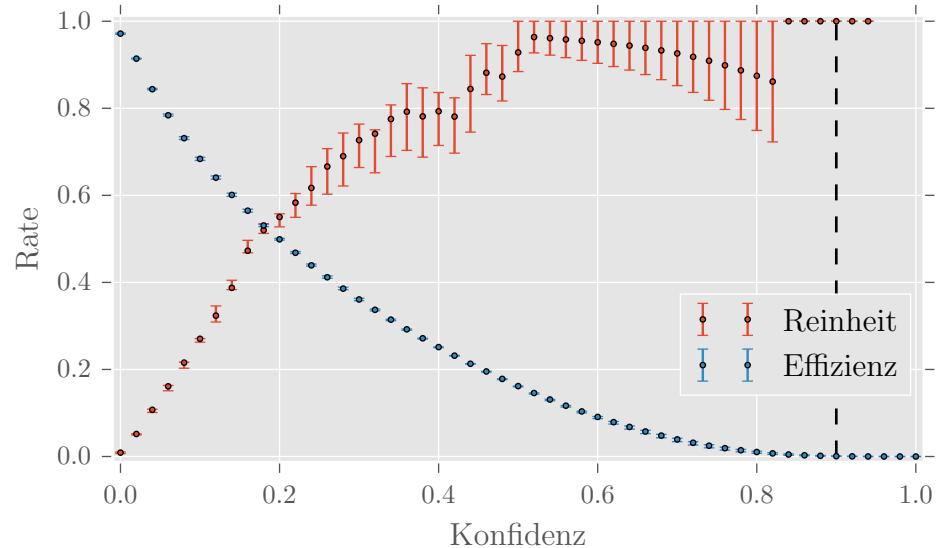


Abbildung A.6: Aus der Neutrino-separation mit dem EXT-Veto. Aufgetragen sind die gemittelten Qualitätsparameter Reinheit (rot) und Effizienz (blau) mit ihren Standardabweichungen gegen die Konfidenz.

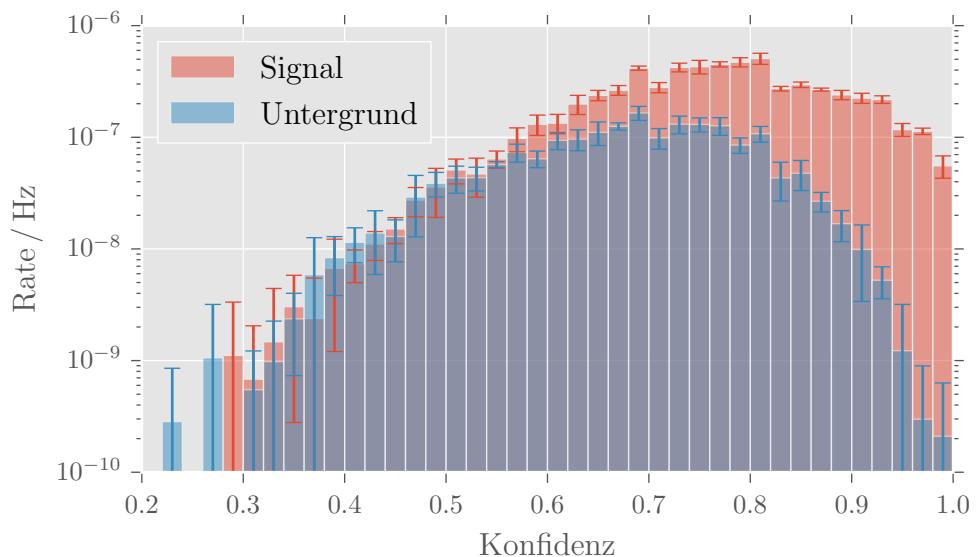


Abbildung A.7: Aus der Myon-Neutrino-separation mit dem EXT-Veto. Histogramme der Konfidenzverteilungen der gemittelten Ereignisraten von Myonen (Untergrund) und Neutrinos (Signal) in Hertz. Die Standardabweichungen aus der fünf-fachen Kreuzvalidierung sind als Fehlerbalken dargestellt.

A.5 Separationsqualität der EXT-Veto Neutrinoselection

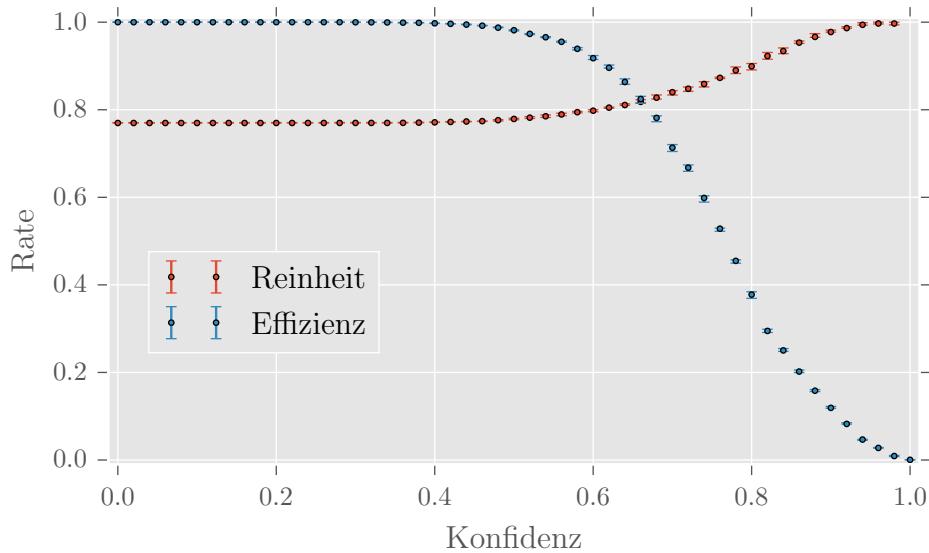


Abbildung A.8: Aus der Myon-Neutrinoseparation mit dem EXT-Veto. Aufgetragen sind die gemittelten Qualitätsparameter Reinheit (rot) und Effizienz (blau) mit ihren Standardabweichungen gegen die Konfidenz.

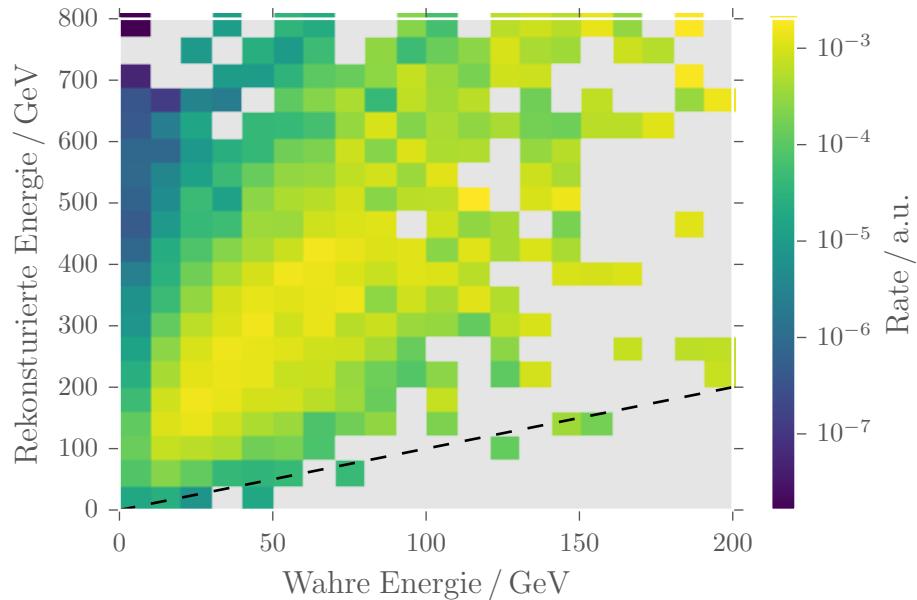


Abbildung A.9: Aufgetragen ist wahre Energie gegen rekonstruierte Energie nach dem Konfidenzschnitt bei 0.9 für das EXT-Veto. Die absoluten Ereignisraten sind farblich kodiert.

Literatur

- [Abb+09] R. Abbasi u. a. „The IceCube data acquisition system: Signal capture, digitization, and timestamping“. In: *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 601.3 (2009), S. 294–316. ISSN: 01689002. DOI: 10.1016/j.nima.2009.01.001. arXiv: 0810.4930.
- [Abb+12] R. Abbasi u. a. „The design and performance of IceCube DeepCore“. In: *Astroparticle Physics* 35.10 (2012), S. 615–624. ISSN: 0927-6505. DOI: <http://dx.doi.org/10.1016/j.astropartphys.2012.01.004>.
- [Abb+13] R. Abbasi u. a. „IceTop: The surface component of IceCube“. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 700 (2013), S. 188–220. ISSN: 0168-9002. DOI: 10.1016/j.nima.2012.10.067.
- [Ach+06] A. Achterberg u. a. „First year performance of the IceCube neutrino telescope“. In: *Astroparticle Physics* 26.3 (2006), S. 155–173. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2006.06.007.
- [Agg15] Charu C. Aggarwal. *Data Mining*. Cham: Springer, 2015. ISBN: 978-3-319-14141-1. DOI: 10.1007/978-3-319-14142-8.
- [And+10] Costas Andreopoulos u. a. „The GENIE neutrino monte carlo generator“. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 614.1 (2010), S. 87–104.
- [Bre01] Leo Breiman. „Random forests“. In: *Machine learning* 45.1 (2001), S. 5–32.
- [Col+08] Andrew Collette u. a. *h5py: Pythonic interface to the HDF5 binary data format*. 2008–. URL: <http://www.h5py.org/> (besucht am 20.08.2016).
- [Cou13] Thierry J.-L. Courvoisier. *High Energy Astrophysics*. Bd. 1. Springer, 2013. ISBN: 9783642309694.
- [Dau+12] Jacob Daughhetee u. a. *DeepCore 2012 Filter Proposal*. Techn. Ber. 2012.
- [De +13] Nicolas De Jay u. a. „MRMRe: An R package for parallelized mRMR ensemble feature selection“. In: *Bioinformatics* 29.18 (2013), S. 2365–2368. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt383.

- [DP05] Chris Ding und Hanchuan Peng. „Minimum redundancy feature selection from microarray gene expression data“. In: *Journal of bioinformatics and computational biology* 3.02 (2005), S. 185–205.
- [Fer49] Enrico Fermi. „On the origin of the cosmic radiation“. In: *Physical Review* 75.8 (1949), S. 1169.
- [Gau+09] Laurent Gautier u. a. *rpy2: Pythonic interface to the HDF5 binary data format*. 2009–. URL: <http://rpy2.bitbucket.org/> (besucht am 20.08.2016).
- [GER16] T.K. Gaisser, R. Engel und E. Resconi. *Cosmic Rays and Particle Physics*. Cambridge University Press, 2016. ISBN: 9781316598436.
- [Hal+09] Mark Hall u. a. „The WEKA data mining software: an update“. In: *ACM SIGKDD explorations newsletter* 11.1 (2009), S. 10–18.
- [Hec+98] Dieter Heck u. a. *CORSIKA: A Monte Carlo code to simulate extensive air showers*. Techn. Ber. 1998.
- [Hör04] Jörg R Hörandel. „Models of the knee in the energy spectrum of cosmic rays“. In: *Astroparticle Physics* 21.3 (2004), S. 241–265.
- [Hun07] J. D. Hunter. „Matplotlib: A 2D graphics environment“. In: *Computing In Science & Engineering* 9.3 (2007), S. 90–95.
- [Jac12] Paul Jaccard. „The distribution of the flora in the alpine zone“. In: *New phytologist* 11.2 (1912), S. 37–50.
- [Koh+95] Ron Kohavi u. a. „A study of cross-validation and bootstrap for accuracy estimation and model selection“. In: *Ijcai*. Bd. 14. 2. 1995, S. 1137–1145.
- [McK+] Wes McKinney u. a. *pandas: Python Data Analysis Library*. URL: <http://pandas.pydata.org/> (besucht am 20.08.2016).
- [Mie+06] I. Mierswa u. a. „YALE: Rapid Prototyping for Complex Data Mining Tasks“. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)* (Aug. 2006), S. 935–940. URL: http://rapid-i.com/component?option=com_docman&task=doc_download&gid=25&Itemid=62.
- [Oli+] T. Oliphant u. a. *Numpy*. URL: <http://www.numpy.org/> (besucht am 20.08.2016).
- [Oli+14] K.A. Olive u. a. „Review of Particle Physics“. In: *Chin.Phys.* C38 (2014), S. 090001. DOI: 10.1088/1674-1137/38/9/090001.
- [Qui86] J. Ross Quinlan. „Induction of decision trees“. In: *Machine learning* 1.1 (1986), S. 81–106.

- [R D08] R Development Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2008. URL: <http://www.R-project.org>.
- [RB91] Guido van Rossum und Jelke de Boer. „Interactively testing remote servers using the Python programming language“. In: *CWI Quarterly* 4.4 (1991). Current state: <https://www.python.org/>, S. 283–303.
- [RMS12] Tim Ruhe, Katharina Morik und Benjamin Schowe. „Data Mining on Ice“. In: *Astrostatistics and Data Mining*. Hrsg. von Manuel Luis Sarro u. a. New York, NY: Springer New York, 2012, S. 223–231. ISBN: 978-1-4614-3323-1. DOI: 10.1007/978-1-4614-3323-1_23.
- [Sch10] Benjamin Schowe. *RapidMiner Feature Selection Extension*. 2010. URL: <http://sourceforge.net/projects/rm-featselext/> (besucht am 20.08.2016).
- [TheNN] The HDF Group. *Hierarchical Data Format, version 5*. 1997-NNNN. URL: <http://www.hdfgroup.org/HDF5/> (besucht am 20.08.2016).
- [WBM98] B. Wiebel-Sooth, P. L. Biermann und H. Meyer. „Cosmic rays. VII. Individual element spectra: prediction and data“. In: *Astronomy and Astrophysics* 330 (Feb. 1998), S. 389–398. eprint: arXiv:astro-ph/9709253. URL: <http://adsabs.harvard.edu/full/1998A&A...330..389W> (besucht am 20.08.2016).

Eidesstattliche Versicherung

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem Titel „IceCube DeepCore Niederenergie Veto-Studie“ selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum

Unterschrift

Belehrung

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50 000 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden (§ 63 Abs. 5 Hochschulgesetz –HG–).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z. B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen.

Ort, Datum

Unterschrift