Datenanalyse mit IceCube-Monte-Carlo-Simulationsdaten

FP2 Teilchen - Astroteilchenphysik

15. April 2021

1 Einleitung

In diesem Versuch wird eine Neutrinoselektion anhand von Monte-Carlo-Simulationsdaten des IceCube-Experiments durchgeführt. Zunächst werden die Daten für die Analyse vorbereitet. Anschließend wird eine Attributsauswahl durchgeführt, um daraufhin mittels maschinellen Lernern eine Separation durchzuführen.

1.1 Grundlagen der Astroteilchenphysik

Informationen über astrophysikalische Objekte werden seit mindestens fünftausend Jahren mithilfe der optischen Astronomie gesammelt. Die Entwicklung von Teleskopen ab dem 17. Jahrhundert und besonders die Erweiterung auf andere Wellenlängenbereiche des elektromagnetischen Spektrums im 20. Jahrhundert mit der Entwicklung der Radioastronomie sowie der Röntgen-, Infrarot- und Ultraviolettastronomie seit Beginn des Raumfahrtzeitalters haben die Kenntnisse über das Universum vervielfacht.

Seit der Entdeckung der Höhenstrahlung durch Viktor Heß im Jahre 1912 ist bekannt, dass die Erde nicht nur von Photonen, sondern auch von anderen Teilchen getroffen wird. Die Höhenstrahlung besteht aus den Sekundärteilchen von Wechselwirkungen der geladenen kosmischen Strahlung. Sie war bis zur Entwicklung der Teilchenbeschleuniger in den 1950-er Jahren die einzige Quelle hochenergetischer Teilchenstrahlen. Die geladene kosmische Strahlung besteht hauptsächlich aus Protonen sowie in unterschiedlichen Anteilen aus Helium und schwereren Kernen. Die genaue Komposition hängt vom Energiebereich ab. Das Energiespektrum der geladenen kosmischen Strahlung erstreckt sich bis zu Energien im Bereich von $10^{20}\,\mathrm{eV}$ und folgt annähernd einem Potenzgesetz

$$\frac{\mathrm{d}\Phi}{\mathrm{d}E} = \Phi_0 E^{\gamma} \tag{1}$$

mit einem spektralen Index γ von etwa $-2{,}7$. Da Protonen und Kerne geladene Teilchen sind, werden sie von galaktischen und extragalaktischen Magnetfeldern abgelenkt. Die Quellen der kosmischen Strahlung sind daher bislang unbekannt.

1.2 Atmosphärische und astrophysikalische Leptonen

Die in IceCube untersuchten Myonen und Neutrinos stammen aus der Atmosphäre oder aus astrophysikalischen Quellen. Die atmosphärischen Myonen und Neutrinos werden unterteilt in konventionelle und prompte. Sie unterscheiden sich voneinander und von den astrophysikalischen Neutrinos durch ihre Energieverteilung.

Konventionelle Myonen und Neutrinos entstehen aus Pionen oder Kaonen, die in der Wechselwirkung der geladenen kosmischen Strahlung mit der Atmosphäre erzeugt werden. Aufgrund ihrer vergleichsweise langen Lebensdauer verlieren diese Mesonen Energie, bevor sie in ein Myon und Myonneutrino zerfallen; aus dem Spektrum der geladenen kosmischen Strahlung $\propto E^{-2.7}$ folgt daher ein Spektrum der konventionellen Myonen und Neutrinos $\propto E^{-3.7}$.

Bei hochenergetischen Wechselwirkungen können auch kurzlebige schwere Hadronen wie D-Mesonen oder Λ_c-Baryonen entstehen; die Lebensdauer dieser Teilchen ist so gering,

daß sie ohne nennenswerten Energieverlust zerfallen, so daß die bei (semi-)leptonischen Zerfällen entstehenden prompten Myonen und Neutrinos das Spektrum der geladenen kosmischen Strahlung erben.

Astrophysikalische Quellen, die Hadronen beschleunigen, sollten auch Neutrinos und hochenergetische Photonen emittieren. Als ungeladene Teilchen werden diese von Magnetfeldern nicht abgelenkt, so dass sie auf ihren Ursprung zurückzeigen. Da Neutrinos einen sehr kleinen Wirkungsquerschnitt haben, können sie auch dichte Staubwolken durchdringen und Informationen über das optisch dichte Innere eines astrophysikalischen Objektes liefern, während Gammastrahlung Staubwolken und optisch dichte Medien nicht durchdringen kann.

Die Annahme von Stoßbeschleunigung, wie in [Fer49], führt auf ein Potenzgesetz für den Fluss der astrophysikalischen Neutrinos mit einem spektralen Index von $\gamma \approx -2$.

1.3 Das IceCube-Experiment

IceCube ist ein Experiment zur Detektion hochenergetischer Neutrinos und Myonen. Der Detektor besteht aus dem In-Ice-Array [Ach+06], DeepCore [Abb+12] und IceTop [Abb+13]. Es befindet sich am geographischen Südpol in einer Tiefe von $1450-2450\,\mathrm{m}$ unter der Oberfläche in einer Schicht klaren Eises. An 86 Kabeln befinden sich insgesamt 5160 Photoelektronenvervielfacher, die das schwache Čerenkovlicht detektieren, das entsteht, wenn sich hochenergetische geladene Teilchen schneller als das Licht im jeweiligen Medium bewegen. Die Lichtgeschwindigkeit c im Medium ist gegeben durch $c = c_0/n$ mit dem Brechungsindex n und der Vakuumlichtgeschwindigkeit c_0.

7 Kabel, die dichter mit Photomultipliern besetzt sind und einen geringeren Abstand voneinander haben als die anderen Kabel, bilden die Niederenergieerweiterung DeepCore. Durch die kleineren Abstände und die höhere Effizienz der verwendeten Photomultiplier ist die Energieschwelle für diesen Teil des Detektors niedriger als für das In-Ice-Array. Sie liegt bei etwa $10\,\mathrm{GeV}$ gegenüber etwa $100\,\mathrm{GeV}$.

An der Oberfläche befindet sich das Luftschauer-Experiment IceTop. Die Teilchen eines Luftschauers werden über ihr Čerenkovlicht in lichtdichten Eistanks detektiert. IceTop dient sowohl als Experiment zur Erforschung der kosmischen Strahlung als auch als Veto für das In-Ice-Array.

Neutrinos werden in IceCube durch die Sekundärteilchen aus folgenden Wechselwirkungen

$$\begin{aligned} \nu_\ell(\overline{\nu}_\ell) + A &\to \ell^\mp + X \quad \text{(CC)}, \\ \nu_\ell + A &\to \nu_\ell + X \quad \text{(NC)} \end{aligned}$$

mit Kernen im Eis detektiert. Es findet dabei eine schwache Wechselwirkung über den geladenen (CC) oder neutralen Strom (NC) statt.

Elektronen produzieren einen näherungsweise sphärischen Schauer im Detektor, während Myonen durch ihren deutlich langsameren Energieverlust eine sehr große Reichweite besitzen und eine lange Spur von Licht hervorrufen. Tau-Leptonen haben wegen ihrer kurzen Lebensdauer eine nur geringe Reichweite, bevor sie zerfallen, und haben für nicht

zu hohe Energien eine ähnliche Signatur wie Elektronen. Bei Wechselwirkungen über den neutralen Strom (NC) wird eine Kaskade durch die hadronischen Sekundärteilchen beobachtet, die der durch Elektronen verursachten Kaskade ähnlich ist.

Das von den Myonen selbst erzeugte Čerenkovlicht ist zu schwach für die Detektion. Myonen emittieren in einem Medium e^+e^--Paare und Photonen, die selbst einen Schauer weiterer Photonen und Elektron-Positron-Paare verursachen. Diese Sekundärteilchen erzeugen ebenfalls Čerenkovlicht, das von den Photovervielfachern detektiert werden kann. Näheres dazu kann in [Koe+13] nachgelesen werden.

1.4 Messung von Neutrinos mit IceCube

Die Verwendung von im Detektor wechselwirkenden Ereignissen, sogenannten *starting events*, ist eine Analysetechnik, bei der die äußeren Abschnitte des Detektors als Veto verwendet werden, um atmosphärische Myonen zu verwerfen. Hier ist das effektive Volumen der Analyse kleiner als das gesamte Detektorvolumen. Hier tragen alle Neutrinoflavor in gleicher Weise bei, so daß ein Großteil der Ereignisse Kaskaden aus Wechselwirkungen über den neutralen Strom oder aus Elektron- und Tauneutrinowechselwirkungen über den geladenen Strom sind. Diese Kaskadenereignisse haben eine gute Energie- und eine schlechte Winkelauflösung.

Spuren von durchgehenden Myonen haben eine schlechtere Energie-, aber eine gute Winkelauflösung sowie je nach Energie eine große Reichweite. Das ermöglicht es, das effektive Volumen der Analyse über das Detektorvolumen hinaus zu steigern, indem die Erde als Schild für atmosphärische Myonen verwendet: Myonen, die von unten kommen, müssen aus Neutrinowechselwirkungen stammen. Durch einen Schnitt im rekonstruierten Zenitwinkel könnten so atmosphärische Myonen und Myonneutrinos getrennt werden, wenn die Richtungsrekonstruktion perfekt funktionieren würde. Da die Richtungsrekonstruktion jedoch bei einem kleinen Teil der Ereignisse fehlerhafte Ereignisse liefert, verbessert sich durch einen Schnitt im rekonstruierten Zenitwinkel das Signal-Untergrund-Verhältnis nur von $1{:}10^6$ auf $1{:}10^3$; für die weitere Trennung von fehlrekonstruierten Myonen und Myonneutrinos werden Verfahren des maschinellen Lernens verwendet. Die Signal-Untergrund-Trennung für diesen Analyseansatz//ist das Ziel dieses Versuchs.

2 Signal-Untergrund-Trennung

2.1 Vorbereitung der Daten

Bevor die Daten für die Signal-Untergrund-Trennung verwendet werden können, müssen einige vorbereitende Schritte durchgeführt werden. Signal und den Untergrund wurden getrennt voneinander simuliert und weiterverarbeitet. Für diesen Versuch müssen die Dateien eingelesen und sämtliche Attribute entfernt werden, die nur in einem der beiden Datensätze vorkommen. Außerdem darf der Lerner keine Monte-Carlo-Wahrheiten, Eventidentifikationsnummern und Gewichte zum Lernen verwenden. Zudem können auch Werte in den Datensätzen vorkommen, die nicht als Zahl identifiziert werden können, wie NaN oder Inf; das kann beispielsweise der Fall sein, wenn das Attribut das Ergebnis eines Fits ist, und dieser Fit nicht konvergiert ist.

Zudem müssen die Daten mit einem binären Label versehen werden, das Signal und Untergrund voneinander trennt und auf das die Lernmethoden trainiert werden können.

Bei den Datensätzen, die in diesem Versuch verwendet werden, ist bereits ein Attribut namens `label` vorhanden. Monte-Carlo-Wahrheiten sind in Attributen enthalten, deren Name `Weight`, `MC` oder `Corsika` enthält.[1]

2.2 Attributselektion

Ein Datensatz enthält oft viele Attribute mit sehr unterschiedlichem Informationsgehalt in Bezug auf die Zielklasse. Um die Rechenzeit zu verringern ist es oft nützlich, die Attribute auszuwählen, die dem Lerner eine möglichst gute Unterscheidung zwischen Signal und Untergrund erlauben. Dazu betrachten wir zwei Verfahren, die Vorwärtsauswahl (Forward Selection) und die mRMR-Selektion.

2.2.1 Vorwärtsauswahl (Forward Selection)

Die Vorwärtsauswahl fügt einem Modell iterativ die Variable hinzu, die zusammen mit den bereits ausgewählten Variablen die Vorhersagekraft des Modells am meisten erhöht. Die Vorhersagekraft wird oft durch das Ergebnis eines F-Tests bewertet. Bei der Forward Selection ist die Auswahl von dem verwendeten Lerner abhängig, der das Modell erzeugt.

2.2.2 mRMR-Auswahl

Die mRMR-Auswahl (minimum Redundancy, Maximum Relevance) ist nicht von einem Lerner abhängig, sondern betrachtet die Wahrscheinlichkeitsverteilungen der verfügbaren Variablen. Dabei wird unter anderem der gemeinsame Informationsgehalt zweier Variablen x, y

$$I(x,y) = \int p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \mathrm{d}x \, \mathrm{d}y \tag{2}$$

benutzt, wobei $p(x), p(y), p(x,y)$ die Wahrscheinlichkeitsdichtefunktionen der betreffenden Variablen sind. Die mRMR-Selektion wählt die Variablen iterativ so aus, dass sie

[1] CORSIKA (COsmic Ray SImulation for KAscade) ist ein Simulationsprogramm für Luftschauer.

möglichst stark mit der Zielvariable korreliert sind, aber möglichst wenig untereinander. Als Zielvariable dient für den vorliegenden Fall die Klasse eines Ereignisses.

2.2.3 Stabilitätsanalyse mit dem Jaccard-Index

Die Stabilität der Attributsauswahl gegen statistische Schwankungen im Lerner lässt sich mithilfe des Jaccard-Index untersuchen. Der Jaccard-Index ist ein Maß für die Ähnlichkeit zweier Mengen F_a, F_b und ist definiert über

$$J(F_a, F_b) = \frac{|F_a \cup F_b|}{|F_a \cap F_b|}\,. \tag{3}$$

Die Attributsauswahl wird ℓ-mal auf ℓ Teilmengen des Datensatzes, wie in Gleichung (4) dargestellt, durchgeführt. Der Index erlaubt dann die Ähnlichkeit der verschiedenen Selektionen zu beurteilen. Bei einem Wert nahe an 1,0 ist die Attributsauswahl stabil gegen statistische Schwankungen.

$$\hat{J} = \frac{2}{\ell(\ell-1)} \sum_{i=1}^{\ell} \sum_{j=i+1}^{\ell} J(F_i, F_j) \tag{4}$$

2.3 Multivariate Selektion

In diesem Versuch werden multivariate Lerner zur Klassifikation der Ereignisse verwendet. Gegenüber einfachen geraden Schnitten in den Variablen können diese auch Korrelationen der Variablen untereinander berücksichtigen.

2.3.1 Verschiedene Lernalgorithmen

Der Naive-Bayes-Lerner basiert auf dem Bayes'schen Theorem über bedingte Wahrscheinlichkeiten

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \tag{5}$$

Wenn A die Klassenzugehörigkeit (Signal A oder Hintergrund $\overline{A}$) bezeichnet und B ein Attribut, dann nimmt der Ausdruck

$$Q = \frac{p(A|B)}{p(\overline{A}|B)} = \frac{p(B|A)p(A)}{p(B|\overline{A})p(\overline{A})} \tag{6}$$

dann einen Wert größer als 1 an, wenn das Ereignis mit größerer Wahrscheinlichkeit Signal als Untergrund ist. Wenn mehrere Attribute vorliegen, ergibt sich unter der naiven Annahme, dass die Attribute nur von der Klassenzugehörigkeit abhängen,

$$Q = \frac{p(A|B_1, \ldots, B_n)}{p(\overline{A}|B_1, \ldots, B_n)} = \prod_{i=1}^{n} \frac{p(B_i|A)}{p(B_i|\overline{A})}. \tag{7}$$

Der Random Forest ist ein Klassifikator, der auf dem binären Entscheidungsbaum basiert. Bei einem Entscheidungsbaum wird an jedem Knoten ein Schnitt in einer Variable durchgeführt und die daraus entstandenen Teilmengen werden in den beiden Ästen

des Knotens wiederum durch Schnitte unterteilt, bis entweder eine bestimmte Tiefe des Baumes erreicht ist oder die Blätter nur Ereignisse einer Klasse enthalten. Um die Effekte des Übertrainierens zu minimieren, wird über ein Ensemble unterschiedlicher Entscheidungsbäume gemittelt. Die am häufigsten verwendete Methode um diese Unterschiedlichkeit zu erreichen ist die Implementation nach [Bre01]. Dabei wird jeder Baum auf einer Teilmenge des Trainingsdatensatzes trainiert und die k Attribute, in denen der Baum an jedem Knoten den besten Schnitt sucht, zufällig gewählt. Die Entscheidung c des Random Forest ist dann das arithmetische Mittel über die Entscheidungen P_i der N Einzelbäume

$$c = \frac{1}{N} \sum_{i=1}^{N} P_i, P_i \in \{0, 1\}. \tag{8}$$

c wird auch als Confidence oder Signalness bezeichnet.

Der kNN-Klassifikator (k Nächste Nachbarn) klassifiziert die Klasse eines unbekannten Ereignisses als den Mittelwert der Klassen der k nächsten Nachbarn.

2.3.2 Qualitätsparameter

Um die Güte einer Klassifikation zu bewerten, werden ausgehend von der Zahl der korrekt als Signal (tp, true positive) bzw. Untergrund (tn, true negative) und der Zahl der fälschlich als Signal (fp, false positive) bzw. Untergrund klassifizierten Ereignisse folgende Parameter berechnet:

$$\text{Reinheit} \quad p = \frac{tp}{tp + fp} \tag{9}$$

$$\text{Effizienz} \quad r = \frac{tp}{tp + fn} \tag{10}$$

2.3.3 Kreuzvalidierung

Die Kreuzvalidierung erlaubt die Angabe des Fehlers auf die obigen Qualitätsparameter durch statistische Schwankungen im Trainingsdatensatz. Dazu wird der Trainingsdatensatz in n Teile aufgeteilt. Auf $n - 1$ Teilen wird der Lerner trainiert, das so erstellte Modell wird zur Klassifikation des verbliebenen Teiles verwendet. Dieses Vorgehen wird n Mal wiederholt, so dass jeder der n Teile einmal als Testdatensatz verwendet wurde. Auf diese Weise erhält man n Werte für die obigen Qualitätsparameter und kann einen Fehler auf den Mittelwert dieser Qualitätsparameter angeben.

2.4 Monte-Carlo-Simulationen

Monte-Carlo-Simulationen sind Simulationen eines Experimentes unter Verwendung von Zufallsgeneratoren. Sie werden insbesondere dann verwendet, wenn ein Experiment nicht oder nur sehr aufwändig analytisch beschrieben werden kann, sei es aufgrund seiner Komplexität oder aufgrund des Auftretens stochastischer Effekte.

Eine der wichtigsten Anwendungen von Monte-Carlo-Simulationen ist die Bereitstellung von Pseudodaten, die auf bekannten Eigenschaften von Signal- und Untergrundereignissen

beruhen. Sie erlauben zum Beispiel die Evaluation der Güte von Rekonstruktionsalgorithmen, da die zugrundeliegende Wahrheit bekannt ist. Analog werden Pseudodaten aus Monte-Carlo-Simulationen verwendet um Analysen zur Trennung von Signal und Untergrund zu entwickeln, wie es in diesem Versuch geschieht.

3 Aufgaben

Bereiten Sie die zur Verfügung gestellten Datensätze von Signal und Untergrund für die Analyse vor. Führen sie anschließend eine Attributsauswahl und eine multivariate Separation durch und bewerten sie die Güte der Separation von Signal- und Untergrund-Ereignissen mit den vorgestellten Qualitätsparametern.

Sie haben folgende Wahlmöglichkeiten: Entweder führen Sie die Attributsauswahl mit mehreren Methoden durch und vergleichen das Ergebnis der Separation mit einem maschinellen Lerner für die verschiedenen Sätze von Variablen, oder sie führen eine Attributsauswahl durch und wenden bei der Separation verschiedene Lerner an, deren Ergebnisse sie vergleichen. Im Fall verschiedener Lerner vergleichen Sie einen Naive-Bayes-Klassifikator mit zwei weiteren Lernalgorithmen.

3.1 Geeignete Hilfsmittel

3.1.1 Rapidminer

Rapidminer ist eine Umgebung für maschinelles Lernen Das in den neuesten Versionen verfügbae Tutorial ist als Einführung geeignet.

Für diesen Versuch von besonderer Bedeutung ist die Weka-Extension[Hal+09], die verschiedene Lerner als Operatoren zur Verfügung stellt; die Feature-Selection-Extension[Lee+11], die die mRMR-Feature-Selection enthält; und die Parallel-Processing-Extension, die die Auswertung von Schleifen über z.B. die Anzahl ausgewählter Attribute auf Maschinen mit mehreren Prozessoren beschleunigt.

3.1.2 Python

Auch für die Programmiersprache Python gibt es verschiedene Bibliotheken, die bei der Bearbeitung dieses Versuches nützlich sind. Zu nennen sind insbesondere sklearn, das viele Lerner zur Verfügung stellt; pandas, das für die Vorbereitung der Datensätze von Nutzen ist. Die mRMR-Feature-Selection ist über das R-Paket mRMRe verfügbar.

Literatur

[Abb+12] R. Abbasi et al. „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. URL: http://www.sciencedirect.com/science/article/pii/S0927650512000254.

[Abb+13] 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* 700.0 (2013), S. 188–220. ISSN: 0168-9002. DOI: 10.1016/j.nima.2012.10.067. URL: http://www.sciencedirect.com/science/article/pii/S016890021201217X.

[Ach+06] A. Achterberg et al. „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. URL: http://www.sciencedirect.com/science/article/pii/S0927650506000855.

[Bre01] Leo Breiman. „Random forests“. In: *Machine learning* 45.1 (2001), S. 5–32.

[Fer49] Enrico Fermi. „On the Origin of Cosmic Radiation“. In: *Physical Review* 75 (1949), S. 1169–1174.

[Hal+09] Mark Hall et al. „The WEKA data mining software: an update“. In: *ACM SIGKDD explorations newsletter* 11.1 (2009), S. 10–18.

[Koe+13] J.-H. Koehne et al. „PROPOSAL for muon propagation“. In: *Computer Physics Communications* 184 (9 2013). DOI: 10.1016/j.cpc.2013.04.001. URL: http://dx.doi.org/10.1016/j.cpc.2013.04.001.

[Lee+11] Sangkyun Lee et al. *Feature Selection for High-Dimensional Data with RapidMiner*. Techn. Ber. 1. TU Dortmund University, 2011.