E5b Lehrstuhlversuch - IceCube Data Mining

Kevin Heinicke Markus Stabrin

IceCube Detektor und Physikprogramm

› Detektor für kosmische Strahlung am Südpol seit 2010 (Konstruktion ab 2005)

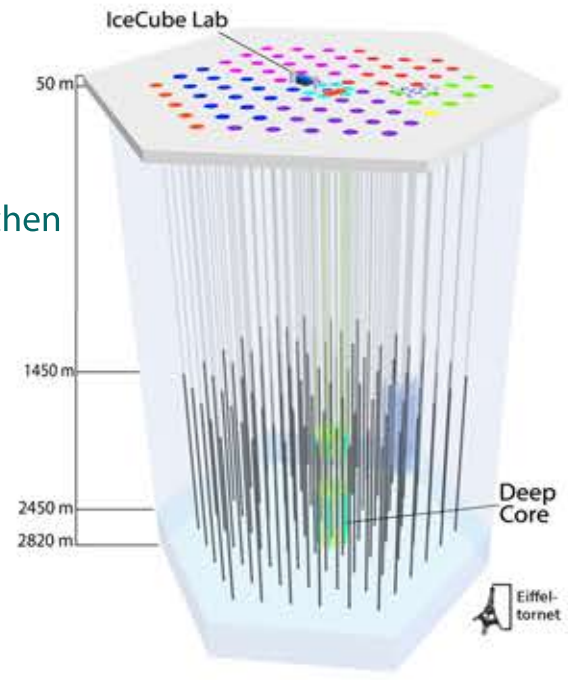
› 5160 Photomultiplier in etwa 1km^3 Eis

› Suche nach hochenergetischen Neutrinos

› Untersuchung kosmischer Strahlung

› Neutrinophysik (PMNS)

› Indirekte Suche nach Dunkler Materie



bub.fysik.su.se/bildarkiv/IceCubeDetector.jpg

Data Mining Datensatz

› 40000 Monte Carlo Ereignisse, 50% Signal, 50% Untergrund

› Insgesamt 280 Features, inklusive **MC Wahrheit**

› Datensatz muss **bereinigt** werden
- › Nur Schnittmenge der Features kann genutzt werden
- › Unphysikalische Werte entfernen (infs, NaNs)
- › Konstante Features besitzen keine Trennkraft
- › MC Wahrheit darf nicht genutzt werden

› **125 Features** nach Bereinigung

Machine Learning Verfahren

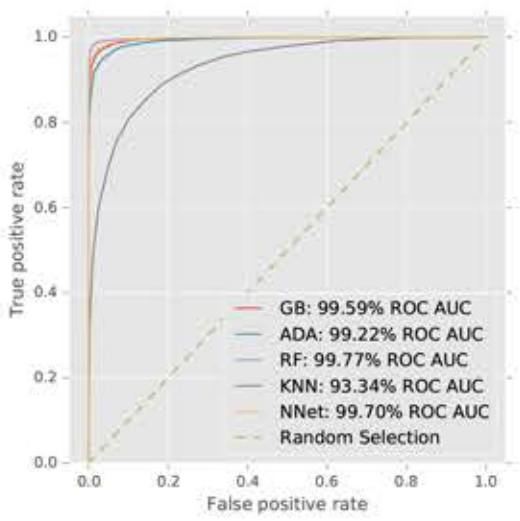
› Test verschiedener ML Verfahren zur Klassifizierung
› Zunächst zufällige Auswahl von Test- und Trainingsdaten im Verhältnis 30/70

› k-Nearest-Neighbours
- › Einfacher Algorithmus, der die Klasse anhand der k (hier k=15) nächsten Nachbarn festlegt
- › anfällig gegenüber dem „Fluch der Dimensionalität“

› Random Forest
- › Ensemble aus Entscheidungsbäumen basierend auf zufällig gewählten Features

› Boosting (AdaBoost, Gradient Boosting)
- › Gewichtete Summe über Entscheidungsbäume, die eine Kostenfunktion minimiert
- › GB: Schwache Features werden sukzessive stärker gewichtet

› Deep Learning
- › Neuronales Netz mit mehr als einer unsichtbaren Schicht (hier: 5)
- › Gewichte der Neuronen werden iterativ verändert, um eine Kostenfunktion zu minimieren



Evaluierung

› Zur Bewertung der Performanz wird die ROC-Curve betrachtet (true positive rate, tpr gg. false positive rate, fpr)

Wahre Klasse	Vorhersage: positive	Vorhersage: negative
negative	false positive	true negative
positive	true positive	false negative

$$\text{tpr} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

$$\text{fpr} = \frac{\text{fp}}{\text{tn} + \text{fp}}$$

$$\text{Reinheit} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{Effizienz} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

› Um