IRFU : Institut de recherche sur les lois fondamentales de l'univers

Saclay

Application des méthodes classiques d'apprentissage supervisé et non supervisé aux données de physique des hautes énergies

Spécialité Physique nucléaire

Niveau d'étude Bac+5

Formation Master 2

Unité d'accueil

Candidature avant le 01/03/2017

Durée 4 mois

Poursuite possible en thèse oui

Contact SABATIE Franck
+33 1 69 08 32 06
franck.sabatie@cea.fr

Résumé

En physique des hautes énergies, les expériences fournissent un grand volume de données dont seulement un faible pourcentage est utile. Le stagiaire appliquera des méthodes connues d'apprentissage automatique, supervisée ou non, afin de filtrer automatiquement ces données.

Sujet détaillé

En physique des hautes énergies, les expériences fournissent un volume important de données. Par exemple, au CERN, chaque seconde où le LHC est en fonctionnement, 25GB de données sont enregistrées. Pour une expérience typique de physique nucléaire de haute énergie à Jefferson Lab, environ 4GB de données sont enregistrées chaque seconde. Une expérience de ce type est prévue en 2017 aux USA, la quantité de données attendues est de 500TB, dont seulement une très petite fraction (moins de 0.01%) contient les événements rares d'intérêt. Il est donc nécessaire de procéder à un filtrage. Aujourd'hui, l'automatisation de ce filtrage est un enjeu car un filtrage sous-optimal induit soit une réduction de la statistique de l'expérience, soit une augmentation significative des erreurs systématiques associées.

Dans ce stage, nous nous proposons d'automatiser ce filtrage par l'apport des méthodes d'apprentissage automatique, en les utilisant sur des données Monte Carlo simulées et bruitées. Pour cela, les méthodes les plus connues dans les domaines de l'apprentissage supervisé ou non-supervisé seront comparées à ce qui se fait dans l'état l'art. On peut citer par exemple les méthodes de clustering k-means et DB-Scan, les réseaux de neurones et les SVM. La question sous-jacente est la caractérisation de la séparabilité des données expérimentales afin d'évoluer notamment vers des algorithmes fournissant une justification en plus de la décision.

Ce sujet de stage s'adresse à des étudiants de niveau M2 et/ou de grande école d'ingénieur. Il pourra être poursuivi par une thèse sous réserve d'obtention d'un financement.

Mots clés

Informatique, intelligence artificielle, physique des particules

Compétences

Informatique, mathématiques, programmation orientée objet

Logiciels

C++, Linux, Windows

Application of supervised and non-supervised machine Learning algorithms to high-energy physics data

Summary

In high energy physics, experiments need to analyze large amount of data, only a small fraction of which is actually useful. The intern will apply known supervised or unsupervised machine learning algorithms for automated data filtering.

Full description

In high energy physics, experiments need to analyze large amount of data. For instance, at CERN, about 25GB of data are recorded every second. For typical high energy nuclear physics experiment at Jefferson Lab, about 4GB of data are recorded every second. Such an experiment is planned by our physics teams in 2017 in the USA. The total amount of data planned for this experiment is 500TB, only a small fraction (less than 0.01%) of which is actually useful later on. It is therefore absolutely crucial to filter the raw data. Nowadays, automated filtering is an issue because sub-optimal filtering will either reduce the amount of final events and/or increase the associated systematic error.

In this internship, we propose to automatize this filtering using automated machine learning algorithms, which will evaluate first on Monte Carlo simulated data with noise. The intern will apply known supervised or unsupervised machine learning algorithms for automated data filtering and compare them with current state-of-the-art analysis techniques. We will test among others the clustering k-means, DB-scan, neural networks and SVM techniques. The underlying question is the characterization of the separability of experimental data in order to evolve to algorithms that include a justification of the decision.

This internship is accessible to M2 and/or engineering schools students. It is potentially extendable towards a PhD, depending on the availability of a PhD grant for the student.

Keywords

Computer science, artificial intelligence, particle physics

Skills

Computer science, maths, object-oriented programming

Softwares

C++, Linux, Windows