Nom et Prénom : AZHARI MOURAD
Date de soutenance : 05/02/2021
Directeur de Thèse : J. ZEROUAOUI

Sujet de Thèse :

Modélisation prédictive à base de Machine Learning: Adaptation et application aux données massives

Résumé :

Les données ne cessent de s'exploser, les technologies continuent à se développer et la prédiction via le Machine Learning remplace l'estimation et l'explication par des méthodes classiques. Dans ce sens, la modélisation prédictive s'étend sur des données provenant de différents champs, partant de marketing, politique, social et passant par la médecine, la biologie et l'écologie , et arrivant au domaine physique particulièrement, la détection des pulsars, la découverte des Particules Exotiques notamment le Boson de Higgs et les Particules Supersymétrique. Dans un premier temps, cette recherche dévoile l'insuffissance des méthodes classiques. Ensuite, elle évoque la révolution des méthodes statistiques avancées, des Méthodes Ensemblistes et l'apprentissage distribué. En outre, elle rapelle les principes et fondements théoriques de quelques approches du Machine Learning. De plus, elle propose une adaptation de quelques méthodes ensemblistes dans le cadre des données déséquilibrées. Enfin, elle avance une nouvelle approche permettant d'analyser des données massives qualifiées de Big Data: Application Programming interface (API) Pyspark. La contrbution principale de cette thèse est la mise en œuvre de l'environnement Pyspark comme une nouvelle technique d'apprentissage distribué qui fait appel aux différentes méthodes de Machine Learning implémentées dans MLlib librairie de l'API Pyspark.

Cette thèse propose la résolution de deux problèmes: Le premier concerne l'adaptation trois méthodes ensemblistes: Bagging, Random Forest, Boosting et ce dans un cadre de données structurées et déséquilibrées comme celles afférentes à la détection des candidats aux pulsars à travers le jeu de données HTRU2. Le second problème s'intéresse à la découverte des Particules Exotiques en utilisant trois jeux de données qualifiés de données massives, provenant du répértoire de Machine Learning public "UCI": "HIGGS", "SUSY" et "HEPMASS". Sous l'API Pyspark, l'expérimentation concerne quatre algorithmes: Deux modèles classiques (Regression Logistique et Arbre de Décision) et deux autres ensemblistes (Random Forest et Gradient Boosted Trees). Les résultats obtenus sont très prometteurs et prouvent que ces algorithmes fonctionnent efficacement et la performance de la prédiction est très bonne en dépassant celle enregistrée dans des travaux antérieurs. Les méthodes de Gradient Boosted Trees et Random Forest agissent mieux que les méthodes de la Régression Logistique et l'Arbre de Décision.

Abstract :

Data continues to explode, technologies keep developing, and prediction via Machine Learning replaces estimation and explanation with classical methods. Thus, predictive modeling extends to different disciplines data as such: marketing, politics, social, medicine, biology, ecology, and physical areas, particularly the detection of pulsar candidates and the discovery of Exotic Particles, especially the Higgs Boson and Supersymmetric Particles. t first, This research reveals the insufficiency of classical methods. Then, it describes the revolution of advanced statistical approaches and Ensemble Methods. Furthermore, it evokes the principles and theoretical bases of some machine learning approaches. Also, this work proposes an adaptation of some ensemble methods in the context of unbalanced dataset. Finally, this study advances a new approach to handle and analyze big datasets with the Pyspark API. The main contribution of this thesis is the application of the Pyspark Environment, as a new distributed learning technique, to the different Machine Learning methods implemented in the MLlib library of the Pyspark API.

This thesis proposes to solve two problems: The first one concerns the adaptation of three ensemble methods as Bagging, Random Forest, and Boosting in the context of unbalanced data such as the detection of pulsar candidates dataset(

HTRU2). The second problem focuses on the discovery of Exotic Particles using three datasets qualified as Big Data, coming from the UCI Machine Learning Repository: "HIGGS", "SUSY" and "HEPMASS". Under the Pyspark API, the experimentation concerns four algorithms: Two classical models (Logistic Regression and Decision Tree) and two ensemble methods (Random Forest and Gradient Boosted Trees). The obtained results are very promising and prove that these algorithms work efficiently and the prediction performance is very good, exceeding that recorded in previous work. The Gradient Boosted Trees and Random Forest methods work better than the Logistic Regression and Decision Tree methods.