AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mr **Manzali Youness**
Soutiendra : **le Samedi 04/01/2025** à **10H00**
*Lieu : **FSDM – Centre Visioconférence***

Une thèse intitulée :

« Improvement and application of machine learning algorithms »

*En vue d'obtenir le **Doctorat***

*FD : **Sciences et Technologies de l'Information et de la Communication***
*Spécialité : **INFORMATIQUE***

Devant le jury composé comme suit :

Nom et prénom	Etablissement	Grade	Qualité
TAIRI Hamid	Faculté des Sciences Dhar El Mahraz, Fès	PES	Président
BENNANI Mohamed Taj	Faculté des Sciences Dhar El Mahraz, Fès	MCH	Rapporteur & Examinateur
GUERSS Fatima-zahra	Institut des Métiers de Sport, Kénitra	MCH	Rapporteur & Examinateur
CHBIHI LOUHDI Mohammed Reda	Faculté des Sciences, Ain Chock Casablanca	MCH	Rapporteur & Examinateur
ZINEDINE Ahmed	Faculté des Sciences Dhar El Mahraz, Fès	PES	Examinateur
RIFFI Jamal	Faculté des Sciences Dhar El Mahraz, Fès	MCH	Examinateur
ELFAR Mohamed	Faculté des Sciences Dhar El Mahraz, Fès	MCH	Directeur de thèse

Résumé :

L'apprentissage automatique est une méthode qui extrait des informations utiles à partir des données. C'est une méthode d'analyse de données qui crée un modèle analytique avec une intervention humaine minimale. La variété et le volume des données disponibles augmentent chaque jour ; cependant, il existe de nombreux algorithmes d'apprentissage automatique, ce qui pose la question de savoir quel algorithme est le plus adapté à chaque type de données. Cette thèse de doctorat introduit de nouvelles approches qui utilisent l'analyse des données pour améliorer les algorithmes d'apprentissage automatique existants.

Cette thèse comprend sept articles de recherche qui contribuent à faire progresser les techniques d'apprentissage automatique. La recherche est organisée en quatre parties, comprenant deux articles abordant divers défis et applications pour améliorer les algorithmes d'apprentissage automatique existants. Le plan de la thèse est le suivant :

Partie I : "Améliorations de l'algorithme des arbres de décision"

Article 1 : "Une nouvelle méthode de pré-élagage des arbres de décision basée sur les probabilités des nœuds."

Cet article propose une nouvelle méthode de pré-élagage basée sur l'algorithme des arbres de décision. La nouvelle approche arrête la création de nœuds d'arbre de décision ayant des probabilités élevées.

Article 2 : "Renforcement des nœuds faibles dans l'algorithme des arbres de décision à l'aide de l'augmentation des données."

Cet article présente un nouvel algorithme EWNDT (Renforcement des nœuds faibles dans l'arbre de décision), qui renforce les nœuds contenant peu d'instances en augmentant leurs données à partir de nœuds d'arbre similaires. Deux approches de similarité entre deux nœuds sont utilisées. Enfin, nous comparons le nouvel algorithme à l'algorithme des arbres de décision.

Partie II : "Techniques de réduction des forêts aléatoires"

Article 3 : "Optimisation du nombre de branches dans une forêt de décision en utilisant les métriques des règles d'association."

Cet article introduit une nouvelle stratégie de réduction des forêts appelée PRM, qui repose sur les métriques des règles d'association. L'approche PRM implique l'extraction de branches de la forêt initiale, l'évaluation du score de chaque branche et la suppression de celles qui sousperforment. Finalement, les branches choisies sont utilisées pour prédire de nouvelles données en agrégeant leurs résultats. Cette stratégie peut être étendue à divers types d'ensembles d'arbres.

Article 4 : "Prédiction de la performance des étudiants en utilisant la forêt aléatoire combinée avec le Naïve Bayes."Dans cet article, nous présentons un nouvel algorithme d'apprentissage automatique qui fusionne la forêt aléatoire avec le Naïve Bayes pour prédire la performance des étudiants. Ce nouvel algorithme est évalué à l'aide de deux ensembles de données d'étudiants distincts et est comparé à sept autres algorithmes d'apprentissage automatique. Les résultats expérimentaux montrent que l'approche proposée atteint un niveau de performance louable.

Partie III : "Algorithmes basés sur KNN (K-nearest neighbors)"

Article 5 : "Un algorithme KNN amélioré basé sur la pondération des caractéristiques et lesrègles d'association."

Cet article présente une nouvelle technique de pondération des caractéristiques développée pour l'algorithme des K-plus proches voisins (KNN). La méthode proposée intègre les capacités des règles d'association et des gains d'information pour attribuer des poids aux caractéristiques individuelles d'un ensemble de données, en tenant compte de leur pertinence pour l'attribut cible. En combinant ces deux techniques puissantes, la nouvelle approche de pondération vise à améliorer les performances de l'algorithme KNN dans diverses tâches d'analyse de données.

Article 6 : "Un algorithme KNN amélioré basé sur les méthodes d'ensemble et la corrélation."

Dans cet article, nous introduisons un algorithme K-plus proches voisins (KNN) amélioré qui utilise KNN comme un apprenant de base au sein d'une méthode d'ensemble, combiné à une sélection de sous-ensembles de caractéristiques basée sur la corrélation. Les résultats expérimentaux démontrent la supériorité de notre algorithme proposé par rapport à d'autres méthodes d'apprentissage automatique.

Partie IV : "Algorithmes de regroupement"

Article 7 : "Une nouvelle méthode de regroupement utilisant les ensembles d'éléments fréquents."

Cet article introduit une nouvelle approche de regroupement centrée autour de la fréquence. Initialement, elle utilise l'algorithme Apriori pour générer des éléments fréquents. Ensuite, elle sélectionne k centres pour maximiser la différence supplémentaire entre ces centres en utilisant une mesure nouvellement élaborée appelée la distance FI. La distance FI combine à la fois la distance euclidienne et la similarité entre les ensembles d'éléments, offrant un critère de regroupement plus robuste. Cette technique de regroupement innovante présente un potentiel pour améliorer l'efficacité et l'efficacité des tâches de regroupement dans divers domaines.

Mots clés : Apprentissage automatique, Algorithmes de classification, techniques d'ensembles, regroupement, exploration de données, l'intelligence artificielle.

IMPROVEMENT AND APPLICATION OF MACHINE LEARNING ALGORITHMS

Abstract :

Machine Learning is a method that extracts useful information from data. It is a method
of data analysis that creates an analytical model with minimal human intervention. The variety and the volume of available data are increasing daily; on the other hand, there are many machine learning algorithms, so the problem is which algorithm is more suitable for each data type. This doctoral thesis introduces new approaches that use data analysis to improve existing machine learning algorithms.

This thesis encompasses seven research papers that contribute to advancing machine learning techniques. The research is organized into four parts, comprising two papers addressing various challenges and applications to improve existing machine-learning algorithms.

The Outline of the thesis is as follows:

Part I: "Decision tree algorithm improvements".

Paper 1: "A new decision tree pre-pruning method based on nodes probabilities."

This paper proposes a new pre-pruning method based on the decision tree algorithm. The new approach stops creating decision tree nodes with high probabilities.

Paper 2: "Enhancing Weak Nodes in Decision Tree Algorithm Using Data Augmentation."

This paper introduces a new algorithm EWNDT (Enhancing weak nodes in the decision tree), which reinforces nodes containing a few instances by increasing its data from similar tree nodes. We have used two approaches for the similarity between two nodes. Finally, we compare the new algorithm with the decision tree algorithm.

Part II: "Random forest pruning techniques"

Paper 3: "Optimizing the number of branches in a decision forest using association rule
metrics."

This paper introduces a novel forest pruning strategy called PRM, which relies on association rules metrics. The PRM approach involves extracting branches from the initial forest, assessing each branch's score, and removing the underperforming ones. Ultimately, the chosen branches are leveraged to predict new data by aggregating their outcomes. Remarkably, this strategy can be extended to various types of tree ensembles.

Paper 4: "Prediction of student performance using random forest combined with na¨ıve
Bayes."

In this paper, we present a novel machine-learning algorithm that merges Random Forest with Naive Bayes to forecast student performance. This new algorithm is evaluated using two distinct student datasets and is compared against seven other machine learning algorithms. The experimental results demonstrate that the proposed approach achieves a commendable level of performance.

Part III: "KNN-based algorithms"Paper 5: "An improved KNN algorithm based on feature weighting and association rules."

This paper presents a novel feature weighting technique developed for the K-nearest neighbors (KNN) algorithm. The proposed method integrates the capabilities of association rules and information gains to assign weights to individual features in a dataset, considering their relevance to the target attribute. By combining these two powerful techniques, the new weighting approach aims to enhance the performance of the KNN algorithm in various data analysis tasks.

Paper 6: "An Improved KNN Algorithm Based on Ensemble Methods and Correlation."

In this paper, we introduce an enhanced K-nearest neighbors (KNN) algorithm that employs KNN as a base learner within an ensemble method, combined with correlation-based feature subset selection. The experimental results demonstrate the superiority of our proposed algorithm over alternative machine learning methods.

Part IV: "clustering algorithms"

Paper 7: "A New Clustering Method Using Frequent Item Sets."

This paper introduces a novel clustering approach centered around frequency. Initially, it employs the Apriori algorithm to generate frequent elements. Subsequently, it selects k centers to maximize the extra difference between these centers using a newly devised measure called the FI-distance. The FI-distance combines both the Euclidean distance and the similarity between item sets, offering a more robust clustering criterion. This innovative clustering technique holds promise for improving the effectiveness and efficiency of clustering tasks in various domains.

Key Words : Machine Learning, classification algorithm, ensemble technique, clustering, datamining, artificial intelligence.