AVIS DE SOUTENANCE DE THESE

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

Mr **SABRI Mohammed**
Soutiendra : **le Vendredi 03/01/2025** à **15H00**
Lieu : ***FSDM – Centre Visioconférence***

Une thèse intitulée :

« Integrating Supervised and Unsupervised Learning for Enhanced Classification in High-Dimensional Data Analysis and Fraud Detection »

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
NFAOUI El Habib	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Président
SABBANE Mohamed	Faculté des Sciences Meknès	PES	Rapporteur
TAIME Abderazzak	Ecole Supérieure de Technologie, Khénifra	MCH	Rapporteur
EL FAZAZY Khalid	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Rapporteur
BALZANELLA Antonio	Università della Campania "Luigi Vanvitelli", Italie	PES	Examinateur
RIFFI Jamal	Faculté des Sciences Dhar EL Mahraz, Fès	MCH	Examinateur
MAHRAZ Mohamed Adnane	Faculté des Sciences Dhar EL Mahraz, Fès	MCH	Examinateur
VERDE Rosanna	Università della Campania "Luigi Vanvitelli", Italie	PES	Directrice de thèse
TAIRI Hamid	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Directeur de thèse
YAHYAOUY Ali	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Co-directeur de thèse

Résumé :

L'apprentissage et l'extraction de connaissances à partir de données à haute dimension sont actuellement des sujets populaires dans les cadres de classification supervisée et non supervisée. Néanmoins, les algorithmes de classification supervisée traditionnels ont souvent du mal à comprendre les motifs complexes dans la représentation des données ou les structures de sous-groupes présentes dans les classes a priori. L'incapacité à intégrer complètement ces motifs peut entraver l'efficacité des classificateurs. Les sous-groupes inclus dans des ensembles de données à grand nombre de dimensions cachent des informations essentielles sur les données d'origine, ce qui peut avoir un effet significatif sur le processus d'apprentissage. Cela reste un défi en apprentissage automatique en raison des nombreuses manières de classer les données en fonction des structure cachés.

Pour combler cette lacune, cette thèse propose une méthodologie d'apprentissage automatique pour révéler les motifs cachés dans divers types d'ensembles de données, afin d'améliorer la performance du classificateur. Le processus commence par l'utilisation d'un algorithme d'apprentissage non supervisé comme étape de prétraitement. Cet algorithme crée des sous-groupes au sein de l'ensemble d'apprentissage en identifiant de nouveaux motifs dans les données. Il y parvient en utilisant des algorithmes de clustering dynamique et de k-means, ainsi qu'une nouvelle objective fonction développée. L'objective fonction vise à minimiser la compacité au sein de chaque sous-groupe et à maximiser la séparation entre les sous-groupes de différentes classes. En outre, ces sous-groupes récemment identifiés ont été utilisés pour entraîner une méthode de classification supervisée, incorporant les centroides de chaque sous-groupe et les poids correspondants, pour classer de nouvelles requêtes. Une version améliorée du classificateur K-Nearest Neighbor a été utilisée pour améliorer la précision des prédictions. Cela a été réalisé en introduisant une nouvelle stratégie intégrée qui inclut une étape de clustering. Cette nouvelle approche améliore non seulement considérablement la précision des prédictions, mais aboutit également à un modèle K-Nearest Neighbor plus simple et compréhensible.

Dans cette thèse, nous apportons une deuxième contribution en étendant notre cadre de classification au domaine de la détection de fraude. Nous développons un algorithme de détection de fraude en utilisant l'algorithme de k-moyennes pour identifier de nouveaux motifs de fraude créés par les fraudeurs. De plus, nous proposons une nouvelle approche de sélection de caractéristiques basée sur le système de poids de clustering associé à chaque sous-groupe. Une variante du classificateur K-Nearest Neighbor est utilisée avec le nouvel espace de caractéristiques sélectionné et les étiquettes augmentées. Notre stratégie globale a prouvé qu'elle améliore significativement les performances de classification sur des ensembles de données réels. Cela démontre son efficacité à réduire les taux d'erreur et à révéler des motifs de fraude complexes.

Mots clés :

Apprentissage supervisé, Apprentissage non supervisé, K-Nearest Neighbor, K-means, Clustering, Détection de fraude, Données de haute dimension, Clustering dynamique, Analyse de données fonctionnelles.

INTEGRATING SUPERVISED AND UNSUPERVISED LEARNING FOR ENHANCED CLASSIFICATION IN HIGH-DIMENSIONAL DATA ANALYSIS AND FRAUD DETECTION

Abstract :
Learning and extracting knowledge from high-dimensional data is currently a popular subject in both supervised and unsupervised classification frameworks. Nevertheless, traditional supervised classifier algorithms frequently struggle to understand the com- plex patterns in data representation or the subgroup structures found in apriori classes. The failure to completely integrate these patterns may hinder the effectiveness of classi- fiers. Subgroups included in datasets with a large number of dimensions hide essential information about the original data, which can have a significant effect on the learning process. This remains a challenge in machine learning due to the many different ways to classify data based on hidden patterns.

In order to address this gap, this thesis proposes a machine learning methodology for showing hidden patterns inside various types of datasets, to enhance the performance of the classifier. The process begins by employing an unsupervised learning algorithm as a preprocessing step. This algorithm creates subgroups within the training set by identifying new patterns in the data. It achieves this by utilizing dynamic clustering and k-means algorithms, along with a newly developed objective function. The objective function aims to minimize the compactness within each subgroup and maximize the separation between subgroups of different classes. Furthermore, these recently identi- fied subgroups were utilized to train a supervised classification method, incorporating the centroids of each subgroup and the corresponding weights, to classify fresh queries. An enhanced version of the K-Nearest Neighbor classifier was utilized to improve the accuracy of predictions. This was achieved by introducing a novel integrated strategy that incorporates a clustering stage. This novel approach not only greatly enhances the accuracy of predictions, but also results in a more simplified and comprehensible K- Nearest Neighbor model.

In this thesis, we make a second contribution by expanding our classification frame-
work to the field of fraud detection. We develop a fraud detection algorithm using the

K-means algorithm to identify new patterns of fraud created by fraudsters. Additionally, we propose a new approach for selecting features based on the clustering weights system associated with each subgroup. A variant of the K-Nearest Neighbor classifier is utilized along with the newly selected feature space and augmented labels. Our com- prehensive strategy has been proven to enhance classification performance significantly on real-world datasets. This demonstrates its effectiveness in lowering error rates and exposing complicated fraud patterns.

Key Words :
Supervised learning, Unsupervised learning, K-Nearest Neighbor, K-means, clustering, Fraud detection, High-dimensional data, Dynamic clustering, Functional data analysis.