REPUBLIQUE DU CAMEROUN
Paix-Travail-Patrie
UNIVERSITE DE YAOUNDE 1
DEPARTEMENT
D'INFORMATIQUE
BP/P.O.Box 812
Yaounde-Cameroun

REPUBLIC OF CAMEROON
Peace-Work-Fatherland
UNIVERSITY OF YAOUNDE 1
COMPUTER SCIENCES
DEPARTMENT
BP/P.O.Box 812
Yaounde-Cameroun

Methode explicable basee sur les auto-encodeurs pour la detection des anomalies dans les flux de donnees

Noms et prnoms : NAKAM YOPDUP MANUELLA KRISTEVA
Matricule : 19M2233
Niveau : Master 2
Spcialit : Sciences de Donnes(DS)
Encadreur : Pr Norbert TSOPZE

Superviseur: Pr Roger ETOUNDI ATSA

Contents

1 Domaine de recherche

Un domaine de recherche est un champ d'tude spcifique qui est trait par les chercheurs. Il est gnralement dfini par un ensemble de questions ou de problmes qui sont tudis, ainsi que par les mthodes et les techniques utilises pour les tudier. il existe plusieurs domaines de recherche mais dans la suite de notre devoir nous nous interesserons au domaine des sciences de donnees.

Les sciences des donnes constituent un domaine interdisciplinaire qui combine des lments des sciences naturelles, des sciences sociales, de l'ingnierie et des mathmatiques. Les sciences des donnes se concentrent sur l'tude, l'analyse et l'interprtation des donnes.

1.1 Bases scientifiques des sciences de donnees

Afin d'exceller dans le domaine des sciences de donnees, des connaissances fondamentales sont ncessaires pour comprendre et pratiquer ce domaine. Entre autre on peut citer:

- **La thorie des probabilits** est utilise pour modliser l'incertitude et la variabilit des donnes.
- **La statistique** est utilise pour analyser les donnes et en tirer des conclusions. Les statistiques jouent un rle fondamental dans les sciences des donnes. Elles fournissent les mthodes et les outils ncessaires pour collecter, analyser et interprter les donnes. Les concepts statistiques tels que la probabilit, l'chantillonnage, l'estimation et les tests d'hypothses sont utiliss pour prendre des dcisions bases sur les donnes.
- **L'apprentissage automatique** est utilis pour dvelopper des modles qui peuvent apprendre partir des donnes. L'apprentissage automatique est une branche des sciences des donnes qui implique la construction de modles et d'algorithmes qui permettent aux ordinateurs d'apprendre partir des donnes. Les bases statistiques et mathmatiques sont utilises pour dvelopper des modles d'apprentissage automatique, tandis que l'informatique fournit les outils et les infrastructures ncessaires pour entraner et dployer ces modles.
- **Informatique** : Les sciences des donnes reposent fortement sur les comptences en informatique. La programmation est essentielle pour manipuler et analyser les donnes grande chelle. Des langages de programmation tels que Python, R et SQL sont couramment utiliss. Les bases de donnes, les systmes de gestion de bases de donnes et les comptences en ingnierie logicielle sont galement ncessaires pour grer efficacement les flux de donnes et dvelopper des systmes d'analyse robustes.

- **Domaine d'application spcifique** : Les sciences des donnes reposent souvent sur des connaissances spcialises dans un domaine d'application spcifique. Par exemple, la bioinformatique, la finance, la mdecine, le marketing ou l'ingnierie peuvent ncessiter des connaissances spcialises pour comprendre les donnes et dvelopper des modles appropris.

1.2 Types de recherche en sciences de donnees

La recherche en sciences des donnes est un domaine en pleine croissance, avec de nombreuses opportunits de recherche dans une grande varit de domaines.La recherche en sciences des donnes peut tre classe de diffrentes manires, en fonction **de l'approche, de l'objectif ou du domaine d'application**.

En fonction de **l'approche**, la recherche en sciences des donnes peut tre divise en deux catgories principales :

- **La recherche fondamentale** vise dvelopper de nouvelles connaissances sur les donnes et les mthodes d'analyse des donnes. Elle est gnralement mene par des universitaires dans des laboratoires de recherche.
- **La recherche applique** vise rsoudre des problmes concrets l'aide des donnes. Elle est gnralement mene par des entreprises, des organisations publiques ou des organismes de recherche applique.

En fonction de **l'objectif**, la recherche en sciences des donnes peut tre divise en deux catgories principales :

- **La recherche exploratoire** vise comprendre les donnes et identifier des tendances ou des schmas. Elle est gnralement utilise pour orienter la recherche future.
- **La recherche confirmatoire** vise tester des hypothses ou valider des modles. Elle est gnralement utilise pour prendre des dcisions ou pour dvelopper des applications.

En fonction du **domaine d'application**, la recherche en sciences des donnes peut tre divise en de nombreuses sous-catgories, telles que :

- **La recherche en sant** utilise les donnes pour amliorer la prvention, le diagnostic et le traitement des maladies.

- **La recherche en finance** utilise les donnes pour prendre des dcisions d'investissement et de gestion des risques.
- **La recherche en marketing** utilise les donnes pour comprendre les comportements des consommateurs et pour dvelopper des stratgies marketing efficaces.
- **La recherche en scurit** utilise les donnes pour dtecter les menaces et les incidents de scurit.

1.3 Methodologie de recherche en sciences de donnees

La mthodologie de recherche en sciences des donnes implique une approche systmatique pour mener des tudes et des projets de recherche bass sur les donnes. Bien que la mthodologie puisse varier en fonction du contexte spcifique et des objectifs de recherche. De maniere generale la mthodologie de recherche en sciences des donnes comprend gnralement les tapes suivantes :

- **Dfinition du problme de recherche:** La premire tape consiste dfinir le problme de recherche. Cela implique de clarifier la question de recherche, de prciser les objectifs de la recherche et d'identifier les donnes et les mthodes d'analyse ncessaires.
- **Collecte des donnes:** La deuxime tape consiste collecter les donnes. Les donnes peuvent tre collectes partir de diverses sources, telles que des bases de donnes, des enqutes, des observations ou des expriences.
- **Nettoyage des donnes :** Une fois que les donnes ont t collectes, elles doivent tre nettoyes. Cela implique de corriger les erreurs, de supprimer les donnes aberrantes et de formater les donnes de manire ce qu'elles soient prtes pour l'analyse.
- **Slection des mthodes et des modles** : Slectionnez les mthodes et les modles d'analyse appropris en fonction de la nature de votre problme de recherche. Cela peut inclure des techniques de statistiques, d'apprentissage automatique, de fouille de donnes, de modlisation prdictive ou d'autres approches analytiques.
- **Analyse des donnes :** La quatrime tape consiste analyser les donnes. L'analyse des donnes peut tre ralise l'aide de diverses mthodes, telles que l'analyse statistique, l'apprentissage automatique ou l'intelligence artificielle.
- **Interprtation des rsultats :** La cinquime tape consiste interprter les rsultats de l'analyse des donnes. Cela implique de tirer des conclusions sur les donnes et de les mettre en relation avec le problme de recherche.

- **Communication des rsultats :** La sixime tape consiste communiquer les rsultats de la recherche. Cela peut tre fait travers des publications scientifiques, des confrences ou des rapports.

La mthodologie de recherche en sciences des donnes est un processus flexible qui peut tre adapt aux besoins spcifiques de chaque projet de recherche.

1.4 Les grands noms du domaine

Le domaine des sciences des donnes est vaste et en constante volution, et de nombreux experts ont contribu de manire significative son dveloppement. Voici quelques-uns des pionniers du domaine des sciences des donnes :

- **Geoffrey Hinton** : Geoffrey Hinton est considr comme l'un des pionniers de l'apprentissage profond (deep learning). Ses travaux sur les rseaux de neurones artificiels ont grandement contribu l'avancement de l'intelligence artificielle et de l'apprentissage automatique.
- **Yann LeCun** : Yann LeCun est un chercheur en intelligence artificielle et directeur de l'AI Research chez Facebook. Il est connu pour ses contributions dans le domaine de la vision par ordinateur et de l'apprentissage profond, notamment pour avoir dvelopp le concept des rseaux de neurones convolutifs (CNN).
- **Andrew Ng** : Andrew Ng est un chercheur et entrepreneur spcialis dans le domaine de l'apprentissage automatique. Il a cofond Google Brain et a t l'un des cofondateurs de Coursera. Ses travaux ont contribu populariser l'apprentissage automatique et rendre les connaissances en sciences des donnes plus accessibles grce des cours en ligne.
- **Judea Pearl** : Judea Pearl est un informaticien et philosophe reconnu pour ses travaux fondamentaux sur le raisonnement causal et les rseaux baysiens. Ses contributions ont permis d'largir les capacits de la modlisation causale dans les sciences des donnes.
- **DJ Patil** : DJ Patil est un data scientist et entrepreneur qui a jou un rle cl dans la popularisation du terme "scientifique des donnes". Il a occup le poste de Chief Data Scientist des tats-Unis sous l'administration Obama et a contribu promouvoir l'utilisation des donnes pour la prise de dcisions dans divers domaines.

- **Fei-Fei Li** : Fei-Fei Li est une chercheuse en intelligence artificielle et vision par ordinateur. Elle a jou un rle important dans le dveloppement de mthodes d'apprentissage profond pour la reconnaissance d'images et a contribu la cration de la base de donnes d'images ImageNet, qui a t largement utilise dans le domaine de la vision par ordinateur.
- **Hadley Wickham** : Hadley Wickham est un statisticien et programmeur renomm, connu pour ses contributions majeures dans le domaine de l'analyse de donnes et de la visualisation. Il a dvelopp des packages populaires tels que ggplot2 et dplyr en langage de programmation R, qui sont largement utiliss par les scientifiques des donnes.

Nous avons aussi des grands noms plus recents tels que:

- **Andrew Ng** est un informaticien amricain qui est considr comme l'un des pionniers de l'apprentissage profond. Il a cofond Coursera, une plateforme d'apprentissage en ligne, et a t nomm directeur de l'intelligence artificielle chez Baidu.
- **Ilya Sutskever** est un informaticien russe qui est galement un pionnier de l'apprentissage profond. Il a cofond OpenAI, une organisation but non lucratif ddie la recherche et au dveloppement de l'intelligence artificielle, et a t directeur de l'intelligence artificielle chez Tesla.
- **Yann LeCun** est un informaticien franais qui est galement un pionnier de l'apprentissage profond. Il est professeur d'informatique l'Universit de New York et directeur du laboratoire de vision et de reconnaissance de formes. Quoc Le est un informaticien amricain qui est un expert en apprentissage automatique et en traitement du langage naturel. Il est directeur de l'ingnierie chez Google AI.
- **Timnit Gebru** est une informaticienne thiopienne-amricaine qui est une experte en apprentissage automatique et en justice sociale. Elle est cofondatrice de Black in AI, une organisation qui vise promouvoir la diversit et l'inclusion dans le domaine de l'intelligence artificielle.
- **Kaggle Community** : Bien que ce ne soit pas une personne spcifique, la communaut Kaggle a jou un rle majeur dans le domaine des sciences des donnes depuis 2010. Kaggle est une plateforme en ligne populaire qui hberge des comptitions de science des donnes et permet aux chercheurs de partager des ensembles de donnes, de collaborer et de repousser les limites de l'apprentissage automatique.

1.5 Conferences et journaux dans le domaine de la data sciences

Il existe plusieurs confrences et journaux renomms dans le domaine des sciences des donnes. Voici quelques exemples parmi les plus influents :

Confrences :

- **Conference on Neural Information Processing Systems (NeurIPS)** : NeurIPS est l'une des principales confrences en apprentissage automatique et en intelligence artificielle. Elle rassemble des chercheurs, des praticiens et des experts du monde entier pour prsenter et discuter des dernires avances dans le domaine.
- **International Conference on Machine Learning (ICML)** : L'ICML est une confrence de premier plan en apprentissage automatique qui se concentre sur la prsentation de travaux de recherche novateurs et sur l'change d'ides entre les scientifiques des donnes.
- **International Conference on Data Mining (ICDM)** : L'ICDM est une confrence de premier plan en fouille de donnes et en extraction de connaissances partir de grandes bases de donnes. Elle rassemble des chercheurs et des praticiens pour partager leurs travaux sur les techniques, les mthodologies et les applications de l'exploration de donnes.
- **International Conference on Very Large Data Bases (VLDB)** : VLDB est une confrence majeure dans le domaine des bases de donnes et de la gestion de donnes grande chelle. Elle offre une plateforme pour prsenter des recherches sur les systmes de gestion de bases de donnes, l'analyse de donnes et les applications lies aux mgadonnes.
- **ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)** : KDD est une confrence de premier plan en science des donnes et en fouille de donnes. Elle explore les dernires avances dans les domaines de la dcouverte de connaissances, de l'apprentissage automatique et de l'exploration de donnes.
- **ACM SIGAI Conference on Artificial Intelligence (AAAI)** : La confrence AAAI sur l'intelligence artificielle (AAAI) est une confrence internationale de premier plan consacre l'avancement de la recherche sur l'intelligence artificielle (IA). Organise chaque anne par l'Association for the Advancement of Artificial Intelligence (AAAI), elle est considre comme l'une des confrences les plus prestigieuses dans le domaine de l'IA. Elle aborde les sujets suivants : Apprentissage automatique (Machine Learning), Apprentissage profond (Deep Learning), Vision par ordinateur

(Computer Vision), Traitement automatique du langage naturel (Natural Language Processing), Raisonnement et planification (Reasoning and Planning), Robotique et agents autonomes (Robotics and Autonomous Agents), Apprentissage par renforcement (Reinforcement Learning) et thique de l'IA (AI Ethics).

Journaux :

- **Journal of Machine Learning Research (JMLR)** : JMLR est un journal en accs libre qui publie des articles de recherche originaux dans le domaine de l'apprentissage automatique et de l'intelligence artificielle. Il est considr comme l'un des journaux les plus prestigieux dans le domaine de l'apprentissage automatique.
- **IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)** : TPAMI est un journal de l'IEEE qui se concentre sur la publication d'articles de recherche sur la vision par ordinateur, la reconnaissance de formes et l'analyse d'images.
- **Data Mining and Knowledge Discovery (DMKD)** : DMKD est un journal qui couvre les aspects thoriques et pratiques de la fouille de donnes, de l'extraction de connaissances et de l'exploration de donnes.
- **Journal of Big Data** : Ce journal se concentre sur la publication d'articles de recherche sur les dfis et les opportunits lis aux donnes massives (big data), y compris les mthodologies, les outils et les applications.
- **ACM Transactions on Knowledge Discovery from Data (TKDD)** : TKDD est un journal qui couvre les aspects de la dcouverte de connaissances partir des donnes, y compris les algorithmes, les modles, les mthodologies et les applications.
- **Knowledge and Information Systems (KAIS)** :est une revue scientifique internationale comit de lecture publie par Springer Nature. Elle est consacre la recherche sur les systmes de connaissance et d'information.
- **L'ACM Transactions on Intelligent Systems and Technology (TIST)** : est une revue scientifique internationale comit de lecture publie par l'Association for Computing Machinery (ACM). Elle est consacre la recherche et la pratique des systmes intelligents et de la technologie.
- **IEEE Transactions on Knowledge and Data Engineering (TKDE)** : est une revue scientifique internationale comit de lecture publie par l'Institute of Electrical and Electronics Engineers (IEEE). Elle est considre comme l'une des principales revues dans le domaine des sciences des donnes et de l'ingnierie des connaissances.

2 Les axes de recherche

2.1 Les differents axes de recherche

En data science, il existe plusieurs axes de recherche qui englobent un large ventail de domaines et de sujets. Voici quelques-uns des principaux axes de recherche :

- **Apprentissage automatique et apprentissage profond** : Cette branche explore les algorithmes et les modles pour permettre aux ordinateurs d'apprendre partir de donnes et de prendre des dcisions sans tre explicitement programms.
- **Analyse de donnes massives (Big Data)** : Ce domaine se concentre sur les mthodes, les outils et les infrastructures pour grer, traiter et analyser des ensembles de donnes massives et complexes.
- **Recherche sur les rseaux neuronaux** : tude approfondie des architectures, des mthodes d'entranement et des applications des rseaux neuronaux, y compris les rseaux convolutifs, rcurrents et d'autres architectures avances.
- **Visualisation de donnes** : Cette discipline cherche dvelopper des mthodes pour reprsenter visuellement des donnes complexes de manire en extraire des informations significatives et faciliter la prise de dcision.
- **Exploration de donnes (Data Mining)** : L'exploration de donnes vise dcouvrir des schmas, des tendances ou des relations significatives au sein de grands ensembles de donnes pour en tirer des informations utiles.
- **Analyse prdictive** : Utilisation de techniques statistiques et d'apprentissage automatique pour prvoir des vnements futurs ou des tendances en se basant sur des donnes historiques.
- **Scurit et confidentialit des donnes** : Cette recherche se concentre sur la protection des donnes sensibles contre les accs non autoriss, les atteintes la vie prive et les cybermenaces.
- **Traitement du langage naturel (NLP)** : Cette branche se focalise sur le dveloppement de modles et d'algorithmes pour permettre aux ordinateurs de comprendre, interprter et gnrer un langage humain.
- **Optimisation et apprentissage par renforcement** : Cette recherche porte sur le dveloppement d'algorithmes permettant de prendre des dcisions squentielles dans des environnements complexes pour atteindre des objectifs spcifiques.

- **Data Science thique et responsable** : Un domaine mergeant qui se concentre sur l'thique de la collecte, de l'utilisation et de la diffusion des donnes, ainsi que sur les implications sociales de l'analyse des donnes.
- **Vision par ordinateur (Computer Vision)** : Il s'agit d'un domaine de recherche qui se concentre sur l'analyse, l'interprtation et la comprhension des images et des vidos par les machines. Cela comprend la reconnaissance d'objets, la dtection de motifs, la segmentation d'images, etc.

Ces axes de recherche en data science ne sont pas exhaustifs et peuvent souvent se chevaucher. De plus, de nouvelles branches mergent constamment mesure que la discipline volue et que de nouveaux dfis se prsentent.

2.2 Axe de recherche choisi

Apres avoir fait les cours fouille de donnees en Master 1 je me suis interessee aux flux de donnees et en particulier a la detection d'anomalies dans les flux de donnees . J'ai donc lu de nombreux articles sur le sujet tel que:

- **Review of Anomaly Detection Algorithms for Data Streams, Tianyuan Lu, Lei Wang * and Xiaoyong Zhao**[**?**]: explore les principaux dfis lis la dtection d'anomalies dans les flux de donnes, tels que la nature dynamique des donnes, la limitation des ressources et la ncessit de dtecter rapidement les anomalies.
- **DeepStream: Autoencoder-Based Stream Temporal Clustering and Anomaly Detection, Shimon Harush, Yair Meidan, Asaf Shabtai**[**?**] : prsente une mthode novatrice appele DeepStream, qui utilise des autoencodeurs pour la dtection d'anomalies et le regroupement temporel dans les flux de donnes.
- **Mthodes parallles pour le traitement des flux de donnes continus, Mme Ge Song**[**?**] : travaille sur les mthodes parallles pour le traitement des flux de donnes continus.
- **Dtection danomalies dans les flux de donnes par structure dindexation et approximation : Application lanalyse en continu des flux de messages du systme dinformation de la SNCF, Lucas Foulon**[**?**] : semble proposer une approche qui combine la structure d'indexation et l'approximation pour dtecter les anomalies dans les flux de messages du systme d'information de la SNCF.

- **Anomaly Detection of Time Series with Smoothness-Inducing Sequential Variational Auto-Encoder**[?] : prsente une approche novatrice pour la dtection d'anomalies dans les sries temporelles. Les auteurs proposent l'utilisation d'un auto-encodeur variationnel squentiel (Sequential Variational Auto-Encoder, SVAE) qui intgre une rgularisation favorisant la rgularit et la continuit des sries temporelles.

C'est donc sur l'article **Anomaly Detection of Time Series with Smoothness-Inducing Sequential Variational Auto-Encoder,Longyuan Li, Junchi Yan, Member, IEEE,, Haiyang Wang, and Yaohui Jin Member, IEEE,** que j'ai decide de m'appuyer avec les conseils de Dr Jiechieu Florentin. Cet artvcle traitant sur les auto-encodeurs; l'axe de recherche choisis est donc : **l'apprentissage automatique**

3 Justification du theme

Le theme qui m'a ete propose est: **Methode explicable basee sur les auto-encodeurs pour la detection des anomalies dans les flux de donnees**

3.1 Definitions de concepts

- **Auto-encodeurs**: sont une forme de rseau de neurones utiliss pour apprendre des reprsentations efficaces des donnes en tentant de reconstruire les entres la sortie. Ils sont composs de deux parties principales : un encodeur et un dcodeur. Ils sont souvent utiliss pour **la compression de donnes, la dtection d'anomalies, la rduction de dimension, la gnration de donnes, etc.**
- **Flux de donnees**: est une squence continue de donnes qui est gnre et traite de manire continue et progressive. Les flux de donnes peuvent provenir de diverses sources, telles que **des capteurs, des transactions, des vnements ou des mdias sociaux**.
- **Anomalies**: se rfrent des observations, des comportements ou des vnements qui diffrent significativement du modle ou du schma habituel des donnes. Elles peuvent tre **le rsultat d'erreurs, de dfauts, de manipulations malveillantes ou simplement de variations inhabituelles dans les donnes**.

3.2 Objectifs du theme

il sera donc question tout au long de mon travail de :

- Trouver un jeu de donnees representant un flux de donnees tels que Apache SPARK :
- En fonction du contexte du jeu de donnees choisis identifier ce qui est considere comme anomalie
- On va appliquer l'algorithme de l'article choisi sur notre jeu de donnees
- On va observer les limites et essayer d'apporter un exemple d'amelioration tout en portant l'interet sur l'explicabilite du modele
- On va appliquer la methode proposee sur notre modele et faire des tests
- Puis on va comparer les resultats

3.3 Les apports du theme

Tout theme de recherche se veut d'un apport soit pour le domaine de la recherche soit pour la vie courante:

- Concernant la recherche, les auto-encodeurs sont generalement des boites noires donc obtenvr une methode explicable est un veritable atout pour faire valider les resultats par un expert et justifier son utilisation.
- D'autres parts les applications de la detection des anomalies dans les flux de donnees peuvent jouer un role important dans de nombreux domaines(sante, securite, ...)

4 Conclusion

Le choix d'un theme de recherche demande de suivre un ensemble d'etapes qui sont autant importantes les unes comme les autres. Le but de cette approche est de se rassurer que le projet respecte les normes de la recherche tels que l'axe de recherche. De plus l'importance et l'utilite du theme doivent etre un indispensable dans le choix de celui-ci.