الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي 8 ماي1945 - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbes

Thesis

To obtain the diploma of **Engineering Degree**

Field of Study: **Computer Science**

Specialization: **Artificial Intelligence and Data Science[AIDS]**

Theme

Implementation and Evaluation of Machine Learning and Deep Learning Methods for Log-Based Anomaly Detection

---

Presented by
Khodja Yousra
Meski Melissa

Defended on: **September, 2025**
In front of the jury composed of

Dr. KHALDI Belkacem	President of the Jury
Dr. BEKKOUCHE Mohammed	Thesis Supervisor
Ms. BENALI Celine	Co-Supervisor
Dr. BEDJAOUI Mohammed	Examiner

Academic Year: 2024/2025

Abstract

Log-based anomaly detection (LAD) has become a critical task for ensuring the reliability and security of large-scale distributed systems. Modern infrastructures, such as Hadoop and Spark, produce massive volumes of logs, making manual analysis infeasible and motivating the use of automated machine learning (ML) and deep learning (DL) solutions. This thesis addresses the LAD problem by designing a complete and reproducible pipeline that combines log parsing, feature engineering, and anomaly detection using a wide spectrum of models. Our contributions are threefold. First, we developed a robust preprocessing framework applied to two benchmark datasets, using session-based grouping and temporal feature extraction for HDFS, and fixed-size time windowing for SPIRIT (which already contains timestamp and temporal fields), while applying both TF–IDF and Word2Vec embeddings to obtain complementary sparse and dense sequence representations. Second, we reimplemented classical ML models (Decision Tree, Random Forest, SVM, Logistic Regression, Isolation Forest, PCA) specifically to integrate temporal features on the HDFS dataset, ensuring fair and consistent comparison. Third, we proposed novel hybrid approaches, including Autoencoder + Clustering, K-Means + IForest, and Incremental PCA, as well as advanced DL models such as a BiLSTM Autoencoder, a BiLSTM-DAGMM hybrid model, and a Variational Autoencoder, achieving improved performance by capturing temporal and structural dependencies in logs. Experiments on HDFS showed that integrating temporal features and hybrid architectures significantly enhances detection accuracy, while on SPIRIT, we demonstrated that shorter fixed-size time windows improve anomaly sensitivity. Finally, we developed a Streamlit-based interactive tool that integrates all components, enabling reproducibility and practical usage. Overall, this work provides a systematic study of LAD, highlights the importance of preprocessing and representation learning (TF–IDF and Word2Vec), and delivers a flexible, extensible framework that can serve as a basis for future research.

Keywords– Log-based Anomaly Detection, HDFS, SPIRIT, TF–IDF, Word2Vec, Temporal Features, Machine Learning, Deep Learning, BiLSTM Autoencoder, DAGMM, Variational Autoencoder, Hybrid Models.

الملخص

أصبحت عملية كشف الشذوذ المعتمدة على سجلات الأنظمة (LAD) مهمة أساسية لضمان موثوقية وأمن الأنظمة الموزعة واسعة النطاق. فالبُنى التحتية الحديثة مثل Hadoop وSpark تنتج أحجاماً هائلة من السجلات، مما يجعل تحليلها اليدوي غير عملي ويدفع إلى اعتماد حلول آلية تعتمد على تقنيات التعلم الآلي (ML) والتعلم العميق .(DL) يتناول هذا البحث مشكلة LAD من خلال تصميم خط معالجة كامل وقابل لإعادة الإنتاج يجمع بين تحليل السجلات، وهندسة الميزات، وكشف الشذوذ باستخدام طيف واسع من النماذج. وتتلخص مساهماتنا في ثلاث نقاط رئيسية: أولاً، قمنا بتطوير إطار قوي للمعالجة المسبقة مطبَّق على مجموعتي بيانات مرجعية، باستخدام تجميع قائم على الجلسات واستخراج الميزات الزمنية لبيانات ،HDFS وتقسيم إلى نوافذ زمنية ثابتة لبيانات SPIRIT (التي تحتوي مسبقاً على حقول زمنية)، مع تطبيق تقنيتي TF--IDF وWord2Vec للحصول على تمثيلات متفرقة وكثيفة متكاملة للتسلسلات. ثانياً، أعدنا تنفيذ النماذج الكلاسيكية للتعلم الآلي (شجرة القرار، الغابة العشوائية، ،SVM الانحدار اللوجستي، Isolation ،Forest وPCA) خصيصاً لدمج الميزات الزمنية في بيانات ،HDFS مما أتاح مقارنة عادلة ومتسقة. ثالثاً، اقترحنا مقاربات هجينة جديدة، تشمل Autoencoder + ،Clustering وK-Means + ،IForest وPCA التزايدي، إضافة إلى نماذج تعلم عميق متقدمة مثل Autoencoder ،BiLSTM والنموذج الهجين ،BiLSTM-DAGMM وAutoencoder التباين ي ،(VAE) والتي حققت أداءً محسّناً بفضل قدرتها على التقاط الاعتماديات الزمنية والبنيوية في السجلات. وقد أظهرت التجارب على بيانات HDFS أن دمج الميزات الزمنية والهياكل الهجينة يحسن بشكل ملحوظ من دقة الكشف، بينما أثبتنا على بيانات SPIRIT أن استخدام نوافذ زمنية أقصر يعزز حساسية كشف الشذوذ. وأخيراً، قمنا بتطوير أداة تفاعلية مبنية على Streamlit تدمج جميع المكونات، مما يتيح إعادة الإنتاج والاستخدام العملي. في المجمل، يقدم هذا العمل دراسة منهجية لمشكلة ،LAD ويسلط الضوء على أهمية المعالجة المسبقة وتمثيل البيانات (TF--IDF وWord2Vec)، ويوفر إطاراً مرناً وقابلاً للتوسعة يمكن أن يشكّل أساساً لأبحاث مستقبلية.

الكلمات المفتاحية -- كشف الشذوذ القائم على السجلات، ،HDFS ،SPIRIT ،TF--IDF ،Word2Vec الميزات الزمنية، التعلم الآلي، التعلم العميق، Autoencoder ،BiLSTM Autoencoder DAGMM التباين ي، النماذج الهجينة.

Résumé

La détection d'anomalies basée sur les journaux (LAD) est devenue une tâche cruciale pour garantir la fiabilité et la sécurité des systèmes distribués à grande échelle. Les infrastructures modernes, telles que Hadoop et Spark, génèrent un volume massif de journaux, rendant leur analyse manuelle impraticable et motivant l'utilisation de solutions automatisées basées sur l'apprentissage automatique (ML) et l'apprentissage profond (DL). Ce mémoire aborde le problème de la LAD en concevant une chaîne de traitement complète et reproductible qui combine le parsing des journaux, l'ingénierie des caractéristiques et la détection d'anomalies à l'aide d'un large éventail de modèles. Nos contributions sont triples. Premièrement, nous avons développé un cadre de prétraitement robuste appliqué à deux jeux de données de référence, en utilisant un regroupement par sessions et l'extraction de caractéristiques temporelles pour HDFS, et un découpage en fenêtres temporelles de taille fixe pour SPIRIT (qui contient déjà des champs temporels), tout en appliquant les représentations TF–IDF et Word2Vec afin d'obtenir des vecteurs de séquences à la fois creux et denses. Deuxièmement, nous avons réimplémenté des modèles classiques de ML (arbre de décision, forêt aléatoire, SVM, régression logistique, isolation forest, PCA) spécifiquement afin d'y intégrer les caractéristiques temporelles sur le jeu de données HDFS, assurant ainsi une comparaison juste et cohérente. Troisièmement, nous avons proposé de nouvelles approches hybrides, incluant Autoencoder + Clustering, K-Means + IForest et PCA incrémental, ainsi que des modèles DL avancés tels qu'un Autoencoder BiLSTM, un modèle hybride BiLSTM-DAGMM et un Autoencoder Variationnel, atteignant de meilleures performances grâce à la capture des dépendances temporelles et structurelles dans les journaux. Les expériences sur HDFS ont montré que l'intégration des caractéristiques temporelles et des architectures hybrides améliore significativement la précision de détection, tandis que sur SPIRIT, nous avons démontré que des fenêtres temporelles plus courtes améliorent la sensibilité aux anomalies. Enfin, nous avons développé un outil interactif basé sur Streamlit intégrant tous les composants, permettant la reproductibilité et l'utilisation pratique. Dans l'ensemble, ce travail propose une étude systématique de la LAD, met en évidence l'importance du prétraitement et de l'apprentissage des représentations (TF–IDF et Word2Vec) et fournit un cadre flexible et extensible pouvant servir de base à de futures recherches.

Mots-clés–Détection d'anomalies, Journaux, HDFS, SPIRIT, TF–IDF, Word2Vec, Caractéristiques temporelles, Apprentissage automatique, Apprentissage profond, BiLSTM Autoencoder, DAGMM, Autoencoder Variationnel, Modèles hybrides.