Détection d'anomalies par partitionnement des séries temporelles multi-variées

Pierre Lotte*, André Péninou**, Olivier Teste**

*IRIT, Université Toulouse 3 - Paul Sabatier, CNRS
118 Route de Narbonne - 31062 Toulouse, France
pierre.lotte@irit.fr,
**IRIT, Université Toulouse 2 - Jean Jaurès, CNRS
118 Route de Narbonne - 31062 Toulouse, France
{andre.peninou, olivier.teste}@irit.fr,

Résumé. Dans cet article, nous proposons une méthode non-supervisée de détection des anomalies dans les séries temporelles multi-variées par partitionnement appelée PARADISE. Cette méthode consiste à créer une partition des variables de la série temporelle étudiée en garantissant la conservation des relations inter-variables servant à identifier les anomalies. Ce partitionnement repose sur un clustering de plusieurs corrélations entre variables. La détection des anomalies est réalisée localement sur chacune des parties. Plusieurs expérimentations effectuées sur des jeux de données synthétiques et réels issus de la littérature, montrent la pertinence de l'approche avec une amélioration significative de la détection d'anomalies.

1 Introduction

Les séries temporelles sont présentes dans de nombreux domaines tels que la santé, la finance, la sécurité des systèmes, l'aéronautique, l'Internet des objets (IOT), les bâtiments connectés (Smart Building) et bien d'autres. Une série temporelle est constituée d'une ou plusieurs variables mesurées au cours du temps. On qualifie une série temporelle d'*uni-variée* ou de *multi-variée* lorsque celle-ci est composée respectivement d'une ou plusieurs variables.

Une des tâches les plus importantes concernant les séries temporelles est la détection d'anomalies qui consiste à identifier les valeurs aberrantes de la série temporelle étudiée. Plusieurs approches ont été proposées dans la littérature scientifique Schmidl et al. (2022).

Les séries temporelles multi-variées sont des ensembles de données complexes constitués de nombreuses variables entre lesquelles peuvent exister des relations inter-variables importantes mais difficiles à exploiter Lejeune et al. (2020). La majorité des approches existantes de détection d'anomalies exploitent les relations temporelles et inter-variables sur l'ensemble des variables qui constituent les séries temporelles multi-variées, sans identifier de sous-ensembles de variables pouvant modéliser localement des phénomènes reliés entre eux, et relativement indépendants des autres variables. La connaissance de ces sous-ensembles de variables pourrait permettre d'identifier et d'exploiter plus facilement les relations inter-variables.

Nous définissons dans cet article une méthode de détection d'anomalies non supervisée par partitionnement de l'ensemble des variables dans les séries temporelles multi-variées. Notre méthode appelée PARADISE (**PAR**tition-based **A**nomaly **D**etection for multivariate t**I**me **SE**ries) identifie dans un premier temps les sous-ensembles de variables fortement corrélées entre elles et faiblement corrélées avec les autres, puis dans un deuxième temps, effectue une détection d'anomalies locale à chaque sous-ensemble (et non à l'ensemble des variables). Cet article vise à montrer l'intérêt de cette démarche sur des jeux de données synthétiques et réels issus de la littérature scientifique.

Nous commençons cet article par un état de l'art de la littérature scientifique portant sur la détection d'anomalies dans les séries temporelles multi-variées. Nous détaillons ensuite la méthode PARADISE proposée. Enfin, nous discutons des expérimentations effectuées dans le but de montrer la pertinence d'une telle approche.

2 État de l'art

Parmi les premières solutions automatiques proposées pour la détection d'anomalies dans les séries temporelles multi-variées, la majorité se base sur l'utilisation de modèles statistiques tels que ARIMA proposé dans Box et al. (2015) ou encore de l'une de ses variantes. Par la suite, les chercheurs se sont penchés sur l'utilisation d'algorithmes d'apprentissage machine. Breunig et al. (2000) proposent par exemple un nouvel algorithme appelé LOF (Local Outlier Factor). Cette méthode se base sur la densité du voisinage de chaque échantillon d'une série temporelle pour en déterminer un degré d'anomalie. Yairi et al. (2001) utilisent l'algorithme K-Means afin de réaliser un clustering permettant d'identifier les motifs récurrents qui peuvent ensuite être liés par des règles d'associations décrivant le comportement normal du système. Liu et al. (2012) proposent l'algorithme Isolation Forest (IF) qui utilise des arbres binaires dans lesquels on tente d'isoler chaque échantillon de la série temporelle à travers un ensemble d'attributs. Pour aller encore plus loin, Cheng et al. (2019) propose de combiner LOF et IF.

Depuis quelques années, la majeure partie des solutions proposées se basent sur des techniques d'apprentissage profond. Les solutions proposées dans Malhotra et al. (2016), Su et al. (2019), Munir et al. (2019), Chen et al. (2020) utilisent une grande variété de modèles tels que des auto-encodeurs basés sur des cellules LSTM, des réseaux de neurones à convolution ou des "echo state networks" (ESN) [1] par exemple. Ces articles se basent tous sur le même principe. Leur entraînement est réalisé uniquement à partir de données ne comportant aucune anomalie. Lorsqu'on leur demande de prédire le futur ou de reconstruire le signal fourni en entrée, ces modèles se retrouveront alors loin de la vérité lorsque les données contiennent des anomalies. Bien que performantes, aucune de ces approches ne prend en compte l'existence de sous-ensembles de variables contenant des relations inter-variables intéressantes.

Dans les approches par apprentissage profond on peut distinguer les méthodes utilisant des réseaux de neurones à graphes (GNN) comme celles de Deng et Hooi (2021) ou encore de Ding et al. (2023). Dans ces approches, on modélise les données sous forme de graphes dans lesquels chaque noeud représente une variable du jeu de données et chaque arête représente une relation inter-variable entre deux noeuds. Bien que cette modélisation induise une séparation

1. ESN : Réseaux de neurones tentant de se rapprocher du modèle d'un cerveau humain en se basant sur des techniques de *reservoir computing*

des variables, la façon dont cette séparation est réalisée ne garantit pas la conservation des relations inter-variables présentes dans le jeu de données utilisé ce qui pourrait provoquer une perte d'informations.

Malgré les gains en performance obtenus par chacune des générations de solution proposées, la détection d'anomalies dans les jeux de données à grande dimensionalité présents dans la littérature scientifique reste difficile pour les algorithmes. Nous suspectons que cette difficulté soit liée à la malédiction de la dimensionalité et à l'absence d'identification de sous-ensembles de variables liées.

3 Description de l'approche PARADISE

3.1 Formalisation du problème

Soit une série temporelle multi-variée notée X composée de d variables et n observations définie par $X = \left(x_{i,j}\right)_{1\leq i\leq n, 1\leq j\leq d}, x_{i,j} \in \mathbb{R}$. On notera $x_{i,\cdot}$ toute observation de X pour l'instant i et $x_{\cdot,j}$ toute variable j de X.

On cherche à affecter à chaque observation $x_{i,\cdot}$ de X un label noté y_i afin d'obtenir le vecteur $Y = (y_i), y_i \in \{0, 1\}, 1 \leq i \leq n$. Les observations considérées comme anormales seront marquées 1 et les observations normales sont marquées 0.

L'approche PARADISE crée une partition de X notée $\mathcal{P}$. Chaque partie notée $X^k = \{x_{\cdot,j_1}, x_{\cdot,j_2}, \dots\}$ est définie comme un ensemble non vide de variables contenues dans X. La partition $\mathcal{P}$ est telle que $X^1 \cup X^2 \cup \cdots \cup X^p = X$ et pour toutes parties X^k et X^l différentes, $X^k \cap X^l = \emptyset$.

On modélise les relations inter-variables par des fonctions notées f. Soient $j_1 \in \{1..d\}$ et $j_2 \in \{1..d\}$ une relation inter-variable existe entre $x_{\cdot,j_1}$ et $x_{\cdot,j_2}$ si $\exists f : \mathbb{R}^n \to \mathbb{R}^n, f(x_{\cdot,j_1}) = x_{\cdot,j_2}$.

On note $\mathcal{F}$ l'ensemble des relations inter-variables, la partition idéale notée $\mathcal{P}^*$ est une partition qui conserve l'ensemble $\mathcal{F}$. Elle respecte les règles suivantes : (i) $\forall f \in \mathcal{F}, \exists X^k \in \mathcal{P} | f(x_{\cdot,j_1}) = x_{\cdot,j_2} \wedge x_{\cdot,j_1} \in X^k \wedge x_{\cdot,j_2} \in X^k$, (ii) Soit l'opérateur $x_{\cdot,j_1} \to x_{\cdot,j_2}$ qui signifie $\exists f \in \mathcal{F} | f(x_{\cdot,j_1}) = x_{\cdot,j_2} \vee f(x_{\cdot,j_2}) = x_{\cdot,j_1}$. On a $\forall x_{\cdot,j_1} \in X^k, \forall x_{\cdot,j_2} \in X^k, \exists \{x_{\cdot,l_1}, \dots, x_{\cdot,l_m}\} \subset X^k | x_{\cdot,j_1} \to x_{\cdot,l_1} \to \cdots \to x_{\cdot,l_m} \to x_{\cdot,j_2}$

3.2 Partitionnement

La première étape mise en place dans PARADISE est une étape de séparation des variables par création d'une partition notée $\widehat{\mathcal{P}^*}$ qui se veut être la plus proche de $\mathcal{P}^*$ ($\mathcal{P}^*$ n'est pas connu dans les jeux de données réels). Pour créer cette partition, nous identifions les relations inter-variables dans les données sans connaissances préalables du système étudié. Pour cela nous utilisons des coefficients de corrélation dans le but de mesurer le degré de relation entre deux variables. D'autres approches de recherche des partitions pourront par la suite être étudiées.

Il existe différents coefficients de corrélation possédant chacun des capacités de détection différentes. Par exemple, le coefficient de Pearson, détecte seulement les relations linéaires. En revanche d'autres coefficients comme ceux de Kendall, Spearman, la corrélation de distance Székely et al. (2007) ou le coefficient ξ Chatterjee (2021) peuvent détecter certaines relations non-linéaires. Pour favoriser la robustesse de notre approche de détection des relations

inter-variables, nous utiliserons ces cinq coefficients. La valeur retenue pour chaque couple de variables est la valeur absolue maximale atteint par un des cinq coefficients.

Nous allons nous baser sur ces coefficients de corrélation entre variables pour déterminer les sous-ensembles de variables possédant des relations inter-variables pertinentes. Pour trouver ces sous-ensembles, notre méthode s'appuie sur l'utilisation d'algorithmes de clustering tels que K-Means ou HDBSCAN. En effet, chaque ligne de la matrice de corrélation donne les coordonnées d'un point dans un espace multi-dimensionnel $(d \times d)$, nous considérons que les points faisant parti d'un sous-ensemble de variables liées seront proches les uns des autres et isolés du reste des points. L'algorithme de clustering utilisé pourra alors identifier ces sous-ensembles.

En utilisant ce principe, PARADISE est capable de proposer une partition $\widehat{\mathcal{P}^*}$ conservant les relations inter-variables au maximum.

3.3 Détection des anomalies par partie

Une fois la partition obtenue, nous pouvons entraîner et exécuter les algorithmes de détection d'anomalies localement sur chacune des parties de manière indépendante. Les algorithmes utilisés acceptent en entrée toute partie $X^k \in \widehat{\mathcal{P}^*}$ préalablement créée et fourniront en sortie un ensemble de scores d'anomalie noté $S^k = \{s_1^k, s_2^k, \ldots, s_n^k\}$ où le score obtenu par chaque observation $x_{i,\cdot}$ de la partie X^k sera noté s_i^k.

À partir des scores locaux affectés aux différentes parties nous devons calculer un score global pour la série temporelle multivariée. Pour ce faire, nous normalisons entre 0 et 1 les scores d'anomalies obtenus par partie de manière indépendante. Cette normalisation a pour but de rendre les scores locaux comparables en terme d'échelle. Ensuite, et puisque nous sommes intéressés principalement par les scores élevés, nous conserverons pour chaque instant de la série temporelle, le score d'anomalie maximal obtenu par toutes les parties.

Puisque nous passons par une première étape de détection locale, nous obtenons, en plus du score d'anomalie global, une indication plus précise sur l'origine de chaque anomalie. En effet, nous savons quelle partie est à l'origine du score de chacune des observations et pouvont donc réduire le nombre de variables candidates ce qui permet d'expliquer en partie l'anomalie.

4 Expérimentations

4.1 Protocole expérimental

Jeux de données synthétiques Nous utilisons des jeux de données synthétiques générés par un outil développé dans le cadre de nos travaux [2]. Les jeux de données générés contiennent un nombre de variables, un taux de contamination et un nombre de sous-ensembles de variables liées configurables. Chaque variable peut être considérée comme une variable de support ou de suivi. Les variables de support sont représentées par des combinaisons linéaires de fonctions sinusoïdales. Leurs équations sont donc décrites par la forme générale $f(x) = \sum_{i=1}^{m} \beta_i osc_i(\alpha_i x)$ avec β_i l'amplitude et α_i la fréquence de la $i^{\text{ème}}$ fonction sinusoïdale osc_i. La fréquence peut également être variable plutôt que fixe.

2. https://gitlab.irit.fr/sig/theses/pierre-lotte/PARADISE

Les variables de suivi sont quant à elles des variables corrélées aux variables de support. Ces variables peuvent être corrélées de manière linéaire, exponentielle ou logarithmique. Leur génération se base sur un système de suites. Si on prend le cas d'une variable de suivi basée sur une corrélation linéaire, les points de la série temporelle générée sont les termes successifs de la suite $u_{m+1} = u_m + sgn\left(f(m+1) - f(m)\right) \times r$ avec r le pas effectué entre chaque point successif, sgn la fonction signe et f la fonction de support. Le principe de construction est identique pour les variables corrélées de façon exponentielle ou logarithmique.

Une fois les données générées, l'outil injecte des anomalies dans certaines variables de chaque sous-ensemble. Les anomalies peuvent être des anomalies de bruit, de fréquence ou encore de suivi des corrélations. Lors de l'injection d'une anomalie, on ne modifie les valeurs que de la variable affectée.

Jeux de données réels Nous utilisons également des jeux de données réels couramment utilisés dans la littérature scientifique. Le premier jeu de données, nommé **SMD**, publié dans Su et al. (2019) contient les relevés de 30 capteurs pour 28 serveurs d'un datacenter. Le deuxième nommé **WADI**, publié dans Ahmed et al. (2017), provient d'un système de distribution de l'eau contenant 93 capteurs. Enfin, le troisième nommé **SWaT**, publié dans Mathur et Tippenhauer (2016), provient d'un système de traitement de l'eau contenant 44 capteurs. Les taux de contaminations de ces jeux de données sont respectivement de 4.21%, 5.77% et 17.37%.

Algorithmes de référence utilisés Nous vérifions l'efficacité de notre méthode avec divers algorithmes. Nous avons retenu IForest (Liu et al. (2012)), LOF (Breunig et al. (2000)), K-Means (Yairi et al. (2001)), DeepANT (Munir et al. (2019)) et HealthESN (Chen et al. (2020)). Le choix de ces algorithmes s'est fait sur deux critères : (i) avoir de la diversité dans les approches utilisées, (ii) retenir les algorithmes parmi ceux ayant le mieux performé dans les expérimentations menées par Schmidl et al. (2022).

4.2 Résultats

QR1 : L'approche PARADISE permet-elle d'améliorer les performances de détection d'anomalies ? Les premières expérimentations réalisées portent sur les jeux de données synthétiques pour lesquels la partition idéale est connue par construction. Les résultats obtenus lors de ces expérimentations sont détaillés dans la table 1. Ces expérimentations ont été menées sur 132 jeux de données comportant entre 5 et 50 variables, 20k observations, entre 2 et 20 parties et à des taux de contamination allant de 0.1% à 10%. Pour effectuer nos expérimentations, nous utiliserons les quatre métriques suivantes : le F1 score (F1), la précision (Pr), le rappel (Ra) et le score ROC (ROC). Le F1 score est obtenu par optimisation du seuil de détection des anomalies grâce à la courbe ROC. Pour PARADISE, le partitionnement obtenu est le meilleur trouvé par clustering avec K-Means et HDBSCAN, optimisé sur le ROC final.

De ces résultats nous pouvons conclure que notre approche permet un gain de performance significatif dans toutes les métriques dans une grande majorité de cas. Le gain obtenu par notre approche dépend des algorithmes utilisés. Certains algorithmes comme HealthESN, DeepANT et K-Means obtiennent des gains moyens de 10%, 8% et 7% respectivement entre l'approche classique et le partitionnement idéal pour la métrique ROC. En revanche, les algorithmes basés sur IForest et LOF bénéficient moins du partitionnement. Nous pouvons également observer

	App. Classique				PARADISE				Part. idéal P^*			
Algorithme	F1	Pr	Ra	ROC	F1	Pr	Ra	ROC	F1	Pr	Ra	ROC
DeepANT	.12	.07	.63	.52	.12	.08	.63	.53	**.15**	**.10**	**.65**	**.60**
HealthESN	.11	.07	.57	.57	**.15**	**.10**	**.63**	.66	**.15**	**.10**	.62	**.67**
IForest	.10	.06	**.57**	.51	**.11**	**.07**	.56	**.52**	**.11**	**.07**	.54	**.52**
K-Means	.10	.06	.55	.51	.11	.07	.57	.56	**.13**	**.08**	**.58**	**.58**
LOF	.10	.06	.52	.50	**.11**	**.07**	**.53**	**.53**	**.11**	**.07**	**.53**	**.53**

TAB. 1 – *Résultats obtenus sur les jeux de données synthétiques. Pour les deux métriques, la version la plus performante est en gras et la deuxième est soulignée.*

que la partition proposée par PARADISE n'atteint pas toujours les résultats obtenus par la partition idéale mais dépasse les résultats de l'approche classique. Le partitionnement proposé ne doit donc pas conserver toutes les relations inter-variables. Ceci est sans doute dû au fait que les coefficients de corrélation, bien que pertinents, sont des outils insuffisants.

Nous avons ensuite procédé à l'exécution des mêmes expérimentations sur les jeux de données issus de la littérature scientifique. Les résultats obtenus sont décrits dans la table 2.

	Algorithme	Classique				PARADISE			
		F1	Pr	Ra	ROC	F1	Pr	Ra	ROC
WADI	DeepANT	**.16**	**.09**	**.54**	**.61**	.12	.06	**.54**	.56
	HealthESN	**.16**	**.10**	.65	.42	.12	.06	**.68**	**.63**
	IForest	**.25**	**.15**	**.69**	**.77**	.24	.14	**.69**	.74
	K-Means	.10	.06	.38	.48	**.19**	**.11**	**.72**	**.72**
	LOF	**.11**	**.06**	**.51**	**.51**	.10	**.06**	.48	.50
SWaT	DeepANT	**.34**	**.26**	**.52**	.41	.32	**.26**	.43	**.45**
	HealthESN	**.56**	**.47**	.68	**.60**	.43	.30	**.71**	.46
	IForest	.28	.22	.37	.35	**.30**	**.23**	**.45**	**.36**
	K-Means	**.37**	**.28**	**.52**	**.45**	.27	.21	.35	.38
	LOF	**.36**	**.29**	.49	**.49**	**.36**	.28	.51	**.49**
SMD	DeepANT	**.25**	**.17**	.73	.74	.22	.15	**.74**	**.75**
	HealthESN	**.25**	**.18**	**.77**	**.83**	.23	.15	.76	.82
	IForest	**.23**	**.14**	**.76**	**.83**	**.23**	**.14**	.75	.82
	K-Means	.29	.19	**.81**	**.87**	**.30**	**.20**	**.81**	**.87**
	LOF	**.11**	**.14**	.58	**.67**	**.11**	**.14**	**.59**	**.67**

TAB. 2 – *Résultats obtenus sur les jeux de données réels. L'approche la plus performante pour chaque métrique est en gras.*

Dans cette table, nous pouvons remarquer plusieurs choses. Tout d'abord, l'efficacité de l'approche PARADISE dépend de l'algorithme mais aussi du jeu de données. Par exemple, l'algorithme HealthESN obtient une amélioration importante de ses performances sur WADI mais perd sur SWaT et reste stable sur SMD. Toutefois, l'approche PARADISE n'induit de perte de performance significative que rarement. Les performances atteintes sur les jeux de données réels ne confirment pas totalement les conclusions précédentes. Cela est probablement dû à la complexité plus importante des jeux de données réels pour lesquels les relations inter-

variables sont plus difficiles à détecter. Dans la suite de nos expérimentations, nous écarterons les algorithmes peu sensibles à notre méthode à savoir IForest et LOF.

QR2 : Est-ce que le partitionnement proposé conserve au maximum les relations inter-variables ? L'approche PARADISE doit conserver les relations inter-variables présentes dans les données afin de conserver les informations. Pour vérifier cela, nous avons utilisé la métrique **ARI**. Nous avons étudié les valeurs de cette métrique pour les partitionnements proposés par notre méthode sur les jeux de données de plus de 30 variables comportant entre 4 et 20 sous-ensembles de variables.

Bien que pertinent dans de nombreux cas, notre partitionnement n'est pas assez précis. En effet, la majorité des cas de figures nous donne un score ARI aux alentours de 0.45 avec des pics pouvant aller jusqu'à 0.72. Les jeux de données possédant peu de sous-ensembles de variables semblent être plus difficiles à partitionner, ce qui se traduit par des scores faibles se situant autour de 0.10. Cela explique les différences obtenues par les partitionnements proposé et idéal observées dans la table 1.

5 Conclusion

Nous décrivons dans cet article une nouvelle approche de détection des anomalies par partitionnement dans les séries temporelles multi-variées. Cette approche partitionne les variables des séries temporelles en utilisant des coefficients de corrélation et des algorithmes de clustering. Les parties sont ensuite traitées séparément par les algorithmes de détection d'anomalies qui leur affectent des scores locaux. Ces scores locaux sont utilisés pour calculer un score global pour la série temporelle multi-variée. Après plusieurs expérimentations menées sur des jeux de données synthétiques et issus de la littérature, nous montrons que notre approche permet d'obtenir des gains de performances souvent significatifs peu importe le nombre de parties. Dans la suite de nos travaux nous nous pencherons sur l'amélioration de la méthode de partitionnement afin de détecter plus finement les relations entre variables.

Références

Ahmed, C. M., V. R. Palleti, et A. P. Mathur (2017). WADI : A water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, CySWATER '17, New York, NY, USA, pp. 25–28. Association for Computing Machinery.

Box, G. E. P., G. M. Jenkins, G. C. Reinsel, et G. M. Ljung (2015). *Time Series Analysis : Forecasting and Control*. John Wiley & Sons.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). LOF : Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas Texas USA, pp. 93–104. ACM.

Chatterjee, S. (2021). A New Coefficient of Correlation. *Journal of the American Statistical Association 116*(536), 2009–2022.

Chen, Q., A. Zhang, T. Huang, Q. He, et Y. Song (2020). Imbalanced dataset-based echo state networks for anomaly detection. *Neural Computing and Applications 32*(8), 3685–3694.

Cheng, Z., C. Zou, et J. Dong (2019). Outlier detection using isolation forest and local outlier factor. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, RACS '19, New York, NY, USA, pp. 161–168. Association for Computing Machinery.

Deng, A. et B. Hooi (2021). Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence 35*(5), 4027–4035.

Ding, C., S. Sun, et J. Zhao (2023). MST-GAT : A multimodal spatial–temporal graph attention network for time series anomaly detection. *Information Fusion 89*, 527–536.

Lejeune, C., J. Mothe, A. Soubki, et O. Teste (2020). Shape-based outlier detection in multivariate functional data. *Knowledge-Based Systems 198*, 105960.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2012). Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data 6*(1), 3 :1–3 :39.

Malhotra, P., A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, et G. Shroff (2016). LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection.

Mathur, A. P. et N. O. Tippenhauer (2016). SWaT : A water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, pp. 31–36.

Munir, M., S. A. Siddiqui, A. Dengel, et S. Ahmed (2019). DeepAnT : A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. *IEEE Access 7*, 1991–2005.

Schmidl, S., P. Wenig, et T. Papenbrock (2022). Anomaly detection in time series : A comprehensive evaluation. *Proceedings of the VLDB Endowment 15*(9), 1779–1797.

Su, Y., Y. Zhao, C. Niu, R. Liu, W. Sun, et D. Pei (2019). Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, New York, NY, USA, pp. 2828–2837. Association for Computing Machinery.

Székely, G. J., M. L. Rizzo, et N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6).

Yairi, T., Y. Kato, et K. Hori (2001). Fault Detection by Mining Association Rules from Housekeeping Data. *3*(9).

Summary

In this article, we suggest a novel non-supervised partition based anomaly detection method for anomaly detection in multivariate time series called PARADISE. This methodology creates a partition of the variables of the time series while ensuring that the inter-variable relations remain untouched. This partitioning relies on the clustering of multiple correlation coefficients between variables to identify subsets of variables before executing anomaly detection algorithms locally for each of those subsets. Through multiple experimentations done on both synthetic and real datasets coming from the literature, we show the relevance of our approach with a significant improvement in anomaly detection performance.