Détection d'anomalies dans des Données Mixtes : Évaluation des performances selon les types d'anomalies détectés

Houda GADACHA[1] & Patricia KUBICKI[2]

[1] *UTAC, France, houda.gadacha@utac.com*
[2] *UTAC, France, patricia.kubicki@utac.com*

Résumé.

Dans cette étude, notre objectif consiste à détecter des anomalies dans des données contenant à la fois des variables quantitatives et qualitatives. La plupart des méthodes de détection d'anomalies sont conçues uniquement pour les données quantitatives. Nous proposons d'utiliser une Analyse Factorielle sur Données Mixtes (AFDM) pour extraire des composantes principales numériques. Ces composantes sont ensuite utilisées pour la détection d'anomalies. Nous évaluons la performance de cette approche en utilisant des données simulées comportant différents types d'anomalies : globales, locales, rares et mixtes. Notre objectif est de déterminer les types d'anomalies détectés par chaque modèle.

Mots-clés. Détection d'anomalies, données mixtes, analyse factorielle des données mixtes, types d'anomalies.

Abstract.

In this paper, our aim is to detect anomalies in datasets containing both numerical and categorical attributes. Most outlier detection methods are designed only for numerical data. We propose using a Factor Analysis of Mixed Data (FAMD) to extract principal components. These components are then used for outlier detection. We evaluate the performance of this approach using simulated data containing various anomaly types : global, local, rare and mixed outliers. Our goal is to determine anomaly types detected by each model.

Keywords. Anomaly detection, mixed data, factor analysis on mixed data, outlier types.

1 Introduction

Selon Hawkins (1980), une observation est considérée comme une anomalie par rapport au reste de la population lorsqu'elle est générée par un mécanisme différent. En d'autres termes, une anomalie est une observation qui se distingue significativement des autres données.

La détection d'anomalies est cruciale dans divers domaines tels que la lutte contre la fraude à l'assurance, la détection des maladies et la sécurité informatique. Leur application permet de repérer des comportements suspects et d'améliorer la robustesse des modèles statistiques.

La détection d'anomalies suscite un intérêt marqué dans la littérature comme en témoignent les travaux de Togbe et al (2020), Liu et al (2008), Orair et al (2010), Syarif et al (2012), Breunig et al (2000), Münz et al (2007), et Knorr et al (1999). Bien que de nombreuses méthodes aient été proposées, la plupart se concentrent sur les données numériques. La détection d'anomalies pour les données mixtes, composées à la fois de variables quantitatives et qualitatives, a reçu moins d'attention. En pratique, certaines techniques de pré-traitement comme one hot encoding (OHE) sont souvent utilisées pour convertir les variables qualitatives en forme quantitative, permettant ainsi l'application de méthodes de détection d'anomalies adaptées aux données numériques.

Il existe principalement deux types d'anomalies en données quantitatives : globales et locales. Les anomalies globales sont des observations dont la valeur est significativement différente de celles de la majorité des observations. Ces anomalies se trouvent en dehors de la plage normale de l'ensemble de données. Les anomalies locales quant à elles ne sont pas nécessairement des valeurs extrêmes. Elles se situent dans la plage normale de l'ensemble de données mais sont considérées comme anormales par rapport à leur voisinage Alghushairy et al (2021). Ces deux types d'anomalies nécessitent des méthodes de détection différentes. Dans les données qualitatives, les anomalies sont appréhendées différemment. Certains articles sont focalisés dans la détection des catégories rares ou inattendues par rapport à la distribution normale des données Koufakou et al (2007) et Rokhman et al (2016). D'autres articles cherchent l'incompatibilité dans les combinaisons de modalités quand il s'agit de plusieurs variables qualitatives Taha et al (2016). Ainsi, certaines combinaisons de modalités peuvent être considérées comme des anomalies si elles sont inhabituelles ou très peu probables.

Dans cet article, nous poursuivons nos travaux portant sur notre approche novatrice pour la détection d'anomalies dans des jeux de données mixtes Gadacha et al 2023. La méthode repose sur l'application préalable d'une analyse factorielle de données mixtes AFDM suivi de l'analyse des facteurs numériques obtenus pour détecter les anomalies. Cette approche offre la possibilité d'utiliser tous les algorithmes conçus pour les données quantitatives. Nous comparons ces résultats à ceux d'une autre approche classique de pré-traitement de données qui consiste à créer une indicatrice binaire pour chaque modalité. Nous évaluons les performances des modèles et approches dans la détection des différents types d'anomalies.

Ce document est structuré de la manière suivante : dans la section 2 nous présenterons de manière concise l'état de l'art des modèles de détection d'anomalies, en mettant particulièrement l'accent sur ceux qui ont été utilisés dans nos expériences. La section suivante sera consacrée à l'explication de notre approche pour traiter les variables mixtes dans le but de détecter les anomalies. Ensuite, dans la section 4, nous analyserons l'efficacité des différents traitements en termes d'AUC et de types d'anomalies détectés.

2 Etat de l'art

D'après Togbe et al (2020), les méthodes de détection d'anomalies peuvent être regroupées en cinq grandes familles : La première concerne les techniques basées sur les tests statistiques pour identifier les observations qui s'écartent significativement de la distribution normale des

données (ex: Z-score, le test de Grubbs Aggarwal (2017)). La deuxième famille est composée de méthodes de clustering qui regroupent les observations en classes et sont utilisées pour la détection d'anomalies par l'identification des individus trop éloignés (ex: K-means, K-means ++ Arthur et al (2007), Velmurugan et al (2010), K-medoids, Density based spatial clustering of applications with noise DBSCAN Kriegel et al (1996)). La troisième famille concerne les méthodes des plus proches voisins qui identifient l'anomalie en comparant ses caractéristiques à celles de ses voisins les plus proches (ex: K-Nearest Neighbors KNN, Local Outlier Factor LOF Breunig et al (2000)). Dans la quatrième famille, nous trouvons les techniques basées sur le deep learning qui utilisent des réseaux de neurones profonds pour apprendre des représentations des données et détecter les anomalies à partir de ces représentations (ex: Auto-encoders AE et OneClass Neural Networks OCNN Chalapathy et al (2019)). La dernière famille comprend les autres méthodes qui ne rentrent pas directement dans les catégories précédentes comme Isolation Forest Liu et al (2008) et OneClass Support vector machine OC-SVM Schölkopf et al (2000). Chaque famille de méthodes a ses propres avantages, inconvénients et domaines d'application, et le choix de la méthode dépend souvent du contexte spécifique et des caractéristiques des données à traiter.

Dans cet article, nous appliquons 3 méthodes de détection d'anomalies issues de 3 familles différentes. Il s'agit de K-medoids, LOF et IF. K-medoids est une technique de clustering basée sur la distance, similaire à K-means, mais le centroïde de chaque classe est remplacé par un médoïde qui est l'observation la plus centrale de la classe c'est-à-dire l'observation dont la dissimilarité moyenne avec les autres observations de la classe est minimale. Les médoïdes peuvent être utilisés pour détecter des anomalies en calculant la distance entre chaque observation et le médoïde de sa classe. Si la distance est supérieure à un seuil prédéfini, l'observation est considérée comme une anomalie (Munz et al (2007)).

Le deuxième modèle utilisé dans nos expériences est Local Outlier Factor (LOF) qui est une méthode basée sur la densité. Il s'agit de calculer un indice de densité de chaque point et d'en déduire la densité d'atteignabilité locale d'une observation à comparer avec celles de ses plus proches voisins. Le score LOF de l'observation est alors le rapport entre la densité d'atteignabilité moyenne de ses K voisins les plus proches et sa propre densité. Les anomalies sont les observations ayant un LOF > 1. Cela signifie qu'elles sont localisées dans les régions de faible densité.

Isolation Forest (IF) est le troisième modèle utilisé dans nos expériences. C'est une méthode inspirée des forêts aléatoires, utilisée dans la détection d'anomalies. Elle repose sur la construction d'arbres de manière récursive pour isoler les observations. Un score d'anomalie est attribué à chaque observation en sélectionnant aléatoirement une variable et la valeur de coupure. Ce score dépend du nombre de coupures nécessaires pour isoler l'observation, soit la longueur du chemin entre le nœud racine et le nœud terminal. Plus le nombre de coupures est faible plus la probabilité que l'observation soit une anomalie est élevée, et inversement.

3 Détection d'anomalies en données mixtes

La plupart des techniques de détection d'anomalies ne sont compatibles qu'avec des données quantitatives. Pour prendre en compte les données catégorielles, une technique couramment utilisée est One hot encoding (OHE). C'est une méthode de pré-traitement de données qui consiste à transformer les modalités en indicatrices binaires en faisant un tableau disjonctif complet pour les variables qualitatives. Cependant, cette technique peut entraîner une augmentation significative de la dimensionnalité des données, surtout si les variables qualitatives ont de nombreuses modalités. Cela peut poser des problèmes de performance et de temps d'exécution pour les algorithmes de détection d'anomalies, en particulier ceux qui impliquent des calculs complexes comme LOF.

Notre approche (Gadacha et al (2023)) consiste à réaliser une analyse factorielle de données mixtes (AFDM) qui permet d'analyser un jeu de données où les observations sont décrites à la fois par des variables quantitatives et par des variables qualitatives. C'est une technique proposé par Pagès (2004) et qui permet d'obtenir des composantes non corrélées qui sont des combinaisons linéaires des variables initiales. Ces composantes sont utilisées, par la suite, pour la détection d'anomalies à l'aide des algorithmes spécifiques aux données quantitatives.

Comme l'AFDM est une généralisation de l'analyse en composantes principales (ACP), nous nous sommes inspirées des travaux de Joliffe (1986) sur l'ACP et le choix des composantes dans la détection d'anomalies. Il a montré que les premières composantes permettent de détecter les valeurs aberrantes qui respectent la structure de corrélation, alors que, les dernières composantes principales permettent de détecter celles qui ne respectent pas la structure de corrélation entre les variables. Afin de vérifier si cette affirmation est également vraie pour l'AFDM, nous avons décidé d'utiliser dans nos expériences des jeux de données contenant à la fois des anomalies qui respectent et d'autres qui violent la structure de corrélation. L'objectif est de comparer les méthodes et les approches de détection afin de déterminer les types d'anomalies identifiés par les premières et dernières composantes de l'AFDM. Cette analyse nous permettra de sélectionner les composantes appropriées pour détecter chaque type d'anomalies.

4 Expérimentations

Dans cette section, nous présentons les expériences que nous avons réalisées et les résultats.

4.1 Expériences

L'objectif est de simuler des données afin de déterminer les types d'anomalies détectés par chaque modèle. Nous avons simulé N=20000 observations et P=15 variables dont 10 quantitatives et 5 qualitatives. Nous avons simulé deux classes d'observations normales, chacune contenant 45% des données, ainsi que 10% de valeurs aberrantes. Les variables numériques ont d'abord été créées à l'aide d'une distribution normale multivariée à 10 dimensions. Les

variables catégorielles ont été créées à l'aide de la distribution multinomiale, chaque variable contenant une catégorie rare. La fréquence de la modalité rare sur chacune des variables ne dépasse pas 1.5% de la totalité des observations.

Nous avons généré quatre types d'anomalies : globales, locales, rares et mixtes (Tab. 1). Les anomalies globales et locales sont générées à partir des variables quantitatives, tandis que les anomalies rares sont créées à partir des variables qualitatives et sont normales sur les variables numériques. Les anomalies mixtes sont anormales sur les deux types de variables quantitatives et qualitatives.

- Anomalies globales : nous avons créé trois catégories d'anomalies globales : globales avec des valeurs élevées, globales avec des valeurs basses et globales qui ne respectent pas la structure de corrélation. Ces anomalies sont générées en utilisant la "loi des 4 sigma" qui est une extension de la "loi des 3 sigma". Les anomalies sont situées à quatre écarts-types de la moyenne.

- Anomalies locales : nous avons créé d'abord deux classes composées d'observations normales en décentrant la moitié des observations. Les anomalies locales sont situées dans trois régions différentes : la première région contient des anomalies locales par rapport à la première classe, la deuxième région contient des anomalies locales par rapport à la deuxième classe et la dernière région contient des anomalies locales situées au milieu des deux classes. En disposant les anomalies locales de cette manière, nous simulons différents scénarios où les comportements anormaux peuvent se produire par rapport à différentes situations.

- Anomalies rares : ce sont des anomalies qui présentent au moins une modalité rare et qui sont normales sur les données quantitatives.

- Les anomalies mixtes : ces anomalies peuvent être à la fois globales et rares, ou locales et rares.

Observations	Catégories	Proportion	Nombre
données normales		90.12%	18024
Anomalies globales	valeurs elevées	0.80%	160
	valeurs faibles	0.80%	160
	structure de corrélation non respectée	0.90%	180
Anomalies locales		2.50%	498
Anomalies rares		2.38%	476
Anomalies mixtes	globales (elevées) et rares	0.60%	120
	globales (faibles) and rares	0.60%	120
	globales (structure de corrélation non respectée) et rares	0.60%	120
	locales et rares	0.70%	141

Tab. 1 - Caractéristiques des données simulées.

Dans la première expérience, nous évaluons les performances des modèles dans la détection de chaque type d'anomalies séparément. Pour chaque mise en oeuvre, nous nous restreignons aux observations normales et au type d'anomalies concerné. Cela permet de savoir comment chaque modèle se comporte dans la détection des différents types d'anomalies et de comparer leurs performances. Dans la deuxième expérience, nous évaluons les performances des modèles en présence de tous les types d'anomalies sans restriction. Nous utilisons l'ensemble de données comprenant toutes les observations normales ainsi que tous les types d'anomalies. Cette expérience permet d'évaluer la capacité des modèles à détecter les anomalies dans des scénarios plus complexes où plusieurs types d'anomalies peuvent coexister dans les données.

Sur chaque jeu de données, nous avons appliqué le processus de détection d'anomalies décrit ci-après. Tout d'abord, nous avons transformé les données en OHE et nous avons appliqué les 3 algorithmes de détection d'anomalies suivants : IF, LOF et K-medoids. Nous avons effectué une validation croisée avec une recherche en grille des meilleurs hyperparamètres et évaluer les performances sur un autre échantillon test.

Par ailleurs, nous avons appliqué l'AFDM aux données d'apprentissage, de sorte que les observations de l'échantillon de validation et de test ne participent pas à la détermination des composantes principales. Nous avons calculé les composantes pour les observations des échantillons de validation et celui de test considérées comme des observations supplémentaires. Nous avons sélectionné les premières composantes principales sur la base du critère de Kaiser-Guttman (Kaiser (1961)), qui consiste à sélectionner les facteurs pour lesquels la valeur propre est supérieure à 1. Nous avons appliqué les 3 algorithmes de détection d'anomalies d'abord aux premières composantes, puis aux dernières composantes et enfin à toutes les composantes de l'AFDM. Pour chaque modèle, nous avons effectué une validation croisée avec recherche en grille des meilleurs hyperparamètres.

4.2 Résultats

Le tableau 2 résume les résultats des modèles de détection d'anomalies sur les données après OHE et les données après AFDM pour chaque type d'anomalies.

Approches	Méthodes	simulation				
		globale	locale	mixtes	rare	tous types
One hot encoding (OHE)	IF	100%	100%	99.99%	50.95%	90.92%
	LOF	100%	80.36%	99.99%	51.63%	90.10%
	K-medoids	100%	100%	100%	51.78%	90.34%
AFDM (toutes composantes)	IF	99.40%	99.99%	100%	**53.88%**	90.34%
	LOF	100%	100%	100%	43.62%	**91.28%**
	K-medoids	100%	76.81%	100%	45.06%	79.60%
AFDM (premières composantes)	IF	99.95%	87.89%	100%	52.44%	75.08%
	LOF	100%	99.19%	99.99%	45.95%	87.75%
	K-medoids	100%	87.16%	100%	47.63%	72.36%
AFDM (dernières composantes)	IF	68.70%	99.99%	100%	53.35%	86.53%
	LOF	84.18%	99.90%	100%	47.27%	70.03%
	K-medoids	67.60%	67.72%	100%	50.38%	71.83%

Tab. 2 – Tableau comparatif des performances des 3 modèles (IF, LOF et K-medoids) en appliquant les différentes approches pour la détection de chaque type d'anomalies.

Les résultats de nos expériences montrent que les performances des modèles varient en fonction de l'approche utilisée et du type d'anomalies détectés.

1. Les anomalies globales : elles sont bien détectées sur les données encodées, les premières et toutes les composantes de l'AFDM (AUC égale à 1). Néanmoins, elles sont moins bien détectées sur les dernières composantes de l'AFDM, l'AUC varie entre 67.60% et 84.18%).

2. Les anomalies locales : LOF, en tant que modèle dédié à la détection d'anomalies locales, est très efficace sur les premières, les dernières et toutes les composantes de l'AFDM. L'AUC varie entre 99% et 100%. Cependant, il est moins performant sur les données encodées avec un AUC égal à 80.36%. K-medoids semble être performant uniquement sur les données OHE. IF est très performant dans la détection de ce type d'anomalies quelque soit l'approche avec de performances légèrement plus faibles sur les premiers composantes de l'AFDM avec un AUC de 87.89%.

3. Les anomalies mixtes : Elles semblent être facilement détectables par les modèles sur les données après OHE et les composantes de l'AFDM.

4. Les anomalies rares: elles sont les plus difficile à détecter. Le meilleur modèle est IF sur toutes les composantes de l'AFDM avec un AUC de 53.88%. Cela peut s'expliquer par le fait que LOF et K-medoids reposent sur le calcul des distances entre les observations. Lorsque les données sont qualitatives, le calcul peut être plus complexe par rapport aux données quantitatives. Il nécessite souvent des techniques spécifiques, telles que l'utilisation de mesures de similarité appropriées (ex: les coefficients de Jaccard, Russel-Rao et Sokal-Michener). Cette complexité peut affecter les performances des modèles de détection d'anomalies, en particulier ceux qui reposent fortement sur ces calculs.

5. En cas de données contenant différents types d'anomalies : le meilleur modèle est LOF sur toutes les composantes de l'AFDM avec un AUC de 91.28%.

En comparant les performances des méthodes après l'AFDM, nous notons une différence entre les résultats sur les premières composantes principales et les dernières (à l'exception du cas des anomalies mixtes). Les résultats sur les premières composantes principales sont meilleurs dans la détection des anomalies globales. En revanche, les dernières composantes sont meilleurs dans la détection des anomalies rares et locales, à l'exception de K-medoids qui détecte mieux les anomalies sur les premières composantes.

Ces résultats montrent que l'importance des différentes composantes principales : Les premières composantes de l'AFDM expliquent la plus grande partie de la variance totale des données. Ceci les rend efficaces pour détecter les anomalies globales. En revanche, les dernières composantes peuvent contenir des informations plus spécifiques ce qui les rend plus adaptées à la détection d'anomalies rares ou locales.

En synthèse, l'usage de toutes les composantes de l'AFDM permet d'améliorer les performances des modèles en particulier dans la détection des anomalies rares et en présence de tous type d'anomalies.

5 Conclusion

La détection d'anomalies est très importante et utile dans de nombreux domaines notamment la finance, la sécurité, la santé etc. Dans cette étude, nous avons présenté une brève revue de certaines méthodes de détection d'anomalies. Un des inconvénients de la plupart des méthodes est qu'elles sont applicables uniquement sur des variables qualitatives et ne parviennent souvent qu'à identifier des anomalies globales et locales.

Dans cette étude, nous avons présenté notre approche basée sur les composantes de l'AFDM et qui vise à améliorer la détection d'anomalies en particulier dans le cas de données mixtes. Nous avons évalué les performances de cette approche comparée à celle des données en OHE dans la détection de 4 types d'anomalies : globales, locales, rares et mixtes. Les résultats de nos expériences ont montré que les modèles sont très efficaces dans la détection des anomalies mixtes. C'est le cas des anomalies globales quelque soit l'approche à l'exception des dernières composantes de l'AFDM. En revanche, les performances dans la détection des anomalies rares sont globalement faibles, l'usage de notre approche a donné un AUC plus élevé que celui des données après OHE. LOF qui est un modèle dédié à la détection des anomalies locales n'a pas été performant sur les données encodées. Néanmoins, il était efficace sur les composantes de l'AFDM. En réalité, les données peuvent contenir diffèrent types d'anomalies. De ce fait, nous avons réalisé la même expérience sur des données contenant à la fois les 4 types d'anomalies. Les résultats ont montré l'intérêt de notre approche dans la détection des anomalies en présence de plusieurs types d'anomalies. Cependant, des évaluations plus formelles, notamment sur de nouvelles simulations où les anomalies sont moins facilement identifiables, sont nécessaires. Nous poursuivons nos travaux dans ce sens.

Bibliographie

Aggarwal, C. C. (2017). Outlier Analysis (Second Edition ed.). *Springer International Publishing AG*.

Agrawal, S. et Agrawal, J.(2015), Survey on Anomaly Detection using Data Mining Techniques, *Procedia Computer Science*, 708 – 713

Akoglu, L., Tong, H. Vreeken, J. et Faloutsos, C. (2012), Fast and reliable anomaly detection in categorical data. *In CIKM*, pages 415–424.

Angiulli, F. et Fassetti, F. (2009), An efficient algorithm for mining distance-based outliers in very large datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):4:1–57.

Arthur, D. et Vassilvitskii, S. (2007) K-Means++: The Advantages of Careful Seeding Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete algorithms, *Society for Industrial and Applied Mathematics* , Philadelphia, 1027-1035.

Basora, L. , Olive, X., Dubot, D. (2019), Recent Advances in Anomaly Detection Methods Applied to Aviation. Aerospace, *MDPI*, 6 (117), pp.1-27.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). Lof : identifying densitybased local outliers. *In ACM sigmod record* , Volume 29, pp. 93–104. ACM..

Chalapathy, R. et S. Chawla (2019). Deep learning for anomaly detection : A survey, arXiv preprint arXiv:1901.03407.

Chandola, V., Banerjee, A. et Kumar, V. (2009). Anomaly detection : A survey. *ACM computing surveys* 41(3), 15.

Ester, M., Kriegel, H.P., Sander, J. et Xu, X. (1996), A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proceedings of the Second KDD'96 International Conference on Knowledge Discovery and Data Mining* , Portland, OR, USA, pp. 226–231.

Gadacha, H., Kubicki, P., Niang, N. (2023). Détection d'anomalies en présence de données quantitatives et qualitatives. *SFdS2023*, [consulté le 19 février 2024]. Disponible sur : https://drive.google.com/file/d/14qQcZ69m4O3Ez-eunTgzLQzl8CjtpJDF/view

Govaert, G. (2003), Analyse des données, *Lavoisier*, Paris.

Joliffe, I. T. (1986), Principal Component Analysis. *Springer Verlag*.

Kaiser, F. (1961). A note on Guttman's lower bound for the number of common factors, *British Journal of Statistical Psychology*, 14, 1-2.

Knorr, E. M. , Ng, R. T. et Tucakov, V. (1999), Distance-based outliers: algorithms and applications, *The VLDB Journal*, 8(3):237–253.

Koufakou, A., Ortiz, E., Georgiopoulos, M., Anagnostopoulos, G., and Reynolds, K. (2007). A scalable and efficientoutlier detection strategy for categorical data. *In Proceedings of the IEEE International Conference on Tools with ArtificialIntelligence ICTAI*, pages 210–217,

Patras-Peloponnese-Greece.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2008). Isolation forest. *In 2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE.

Munz,G., Li, S. et Carle, G. (2007), Traffic Anomaly Detection Using K-Means Clustering, *GI/ITG Workshop MMBnet*, p.1-8

Orair, G. H., Teixeira, C. H., Wang, Y. et Meira Jr, T. W. (2010), Distance-Based Outlier Detection: Consolidation and Renewed Bearing, *Proceedings of the VLDB Endowment*, Volume 3, p 1469–1480.

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée*, 5–37.

Rokhman, N., Subanar, and Winarko, E. (2016). Improving the performance of outlier detection methods for categorical data by using weighting function. *Journal of Theoretical and Applied Information Technology*, 83:327–336.

Schölkopf, B., R. C. Williamson, A. J. Smola, J. Shawe-Taylor, et J. C. Platt (2000), Support vector method for novelty detection *In Advances in neural information processing systems*, pp. 582–588.

Shyu, M., Sarinnapakorn, K., Kuruppu-Appuhamilage, I., Chen, S., Chang, L. W., et Goldring, T. (2005). Handling nominal features in anomaly intrusion detection problems. *In Proceedings of the International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications*, pages 55–62, Tokyo, Japan.

Syarif I., Prugel-Bennett A. et Wills G. (2012) , Data mining approaches for network intrusion detection from dimensionality reduction to misuse and anomaly detection, *Journal of Information Technology Review*, p. 70-83.

Taha, A. and Hadi, A. S. (2016). Pair-wise association for categorical and mixed attributes. *Information Sciences*, 346:73–89.

Togbe, M., Chabchoub, Y., Boly, A. et Chiky, R. (2020), Étude comparative des méthodes de détection d'anomalies. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances EGC, vol. RNTI-E-36, pp.109-120.

Velmurugan, T. et Santhanam, T. (2010), Computational Complexity between k-Means and k-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points, *Journal of Computer Science*, 6 (3): 363-368.

Wang, Y. (2008). Statistical Techniques for Network Security: Modern Statistically-Based Intrusion Detection and Protection.*IGI Global*, New York, NY, USA.

Wibisono, S., Anwar, M.T , Supriyanto, A., et Amin, I.H.A (2020), Multivariate weather anomaly detection using DBSCAN clustering algorithm, *Journal of Physics: Conference Series* , Volume 1869.

Xu, X., Liu, H. , et Yao, M. (2019), Recent Progress of Anomaly Detection, *Hindawi Complexity*, pages 1-11, January.