Etude comparative des méthodes de détection d'anomalies

Maurras Ulbricht Togbe, Yousra Chabchoub, Aliou Boly, Raja Chiky

Etude comparative des méthodes de détection d'anomalies

Maurras Ulbricht Togbe*, Yousra Chabchoub*
Aliou Boly**, Raja Chiky*

*ISEP, 10 rue de Vanves 92130 ISSY LES MOULINEAUX, France
prenom.nom@isep.fr,
http ://www.lisite.isep.fr
**Université Cheikh Anta Diop de Dakar, BP 5005 Dakar-Fann, Sénégal
prenom.nom@ucad.edu.sn
https ://www.ucad.sn

Résumé. La détection d'anomalies est un problème en plein essor et qui revêt une importance dans plusieurs domaines. A titre d'exemple, la cybercriminalité peut provoquer des pertes économiques considérables et menacer la survie des entreprises. Sécuriser son système d'information est devenu une priorité et un enjeu stratégique pour tous les types d'entreprises. D'autres domaines sont également impactés tels que la santé, les transports, etc. Les solutions de supervision mises en place sont souvent basées sur des algorithmes de détection d'anomalies issus du datamining et du machine learning. Nous présentons dans ce papier un état de l'art complet sur les algorithmes de détection d'anomalies. Nous proposons une classification de ces méthodes en se basant à la fois sur le type de jeux de données (flux, séries temporelles, graphes, etc.), le domaine d'application et l'approche considérée (statistique, classification, clustering, etc.). Nous nous focalisons ensuite sur trois algorithmes : LOF, OC-SVM et Isolation Forest que nous testons sur deux jeux de données différents afin de comparer leurs performances.

1 Introduction

La détection d'anomalies est un volet du datamining qui intéresse de plus en plus de chercheurs actuellement. On trouve dans la littérature plusieurs définitions de l'anomalie souvent appelée outlier. Hawkins (1980) définit un outlier comme une observation qui dévie considérablement du reste des autres observations comme si elle était générée par un processus différent. Quant à Dunning et Friedman (2014), ils affirment que la détection d'anomalie consiste à modéliser ce qui est normal dans le but de découvrir ce qui ne l'est pas. Aggarwal (2017) fait la distinction entre un outlier et une anomalie. Un outlier désigne le bruit et l'anomalie. Le degré d'aberrance permet de différencier les bruits des anomalies.

La détection d'anomalies permet d'améliorer la qualité des données par suppression ou remplacement des données anormales. Dans d'autres cas, les anomalies traduisent un événement et apportent de nouvelles connaissances utiles. Par exemple, la détection d'anomalies peut prévenir un dommage matériel et donc inciter à la maintenance prédictive dans le domaine de l'industrie. Elle trouve son application dans plusieurs autres domaines comme la santé, la cybersécurité, la finance, la prédiction des catastrophes naturelles, et bien d'autres domaines.

Les données existent sous plusieurs formes : les données statiques, les flux de données, les données structurées et non structurées, etc. Chaque type de données est pertinent dans un ou plusieurs domaines. La multitude des types de données et leurs caractéristiques différentes impliquent l'existence de méthodes différentes pour la détection d'anomalies, chacune trouvant son efficacité dans un domaine particulier, avec un objectif donné. Ces méthodes utilisent en général un seuil de décision permettant d'isoler les anomalies en se basant sur les différentes techniques comme la classification, le clustering, la régression, les plus proches voisins et les outils statistiques.

Plusieurs critères peuvent être considérés pour comparer ces méthodes et permettent de choisir la méthode la plus adéquate au contexte : l'implication de l'humain (supervisée, non supervisée, semi-supervisée), la nécessité de faire des hypothèses sur la loi de distribution des données (paramétriques, non-paramétriques, semi-paramétriques), la capacité de traiter des données multivariées et bien d'autres critères. La mesure de la performance de telles méthodes peut s'appuyer sur différents critères comme la précision de la détection (faux positifs, faux négatifs), la rapidité de la détection (temps de réponse) et le passage à l'échelle (par rapport au volume des données ou au débit du flux) entre autres.

Dans cet article, nous proposons d'abord une classification multicritère des méthodes de détection d'anomalies existantes dans la littérature. Puis nous nous focalisons sur trois méthodes : LOF, OC-SVM et Isolation Forest que nous testons sur deux jeux de données différents. Le reste du papier est organisé comme suit : dans la section 2, nous présentons un état de l'art des méthodes de détection d'anomalies. Une classification multicritère de ces méthodes est proposée dans la section 3. Dans la section 4, nous présentons les résultats de notre étude expérimentale comparant les trois méthodes de détection d'anomalies sus-citées. La conclusion est donnée dans la section 5.

2 Etat de l'art

La détection d'anomalies est un sujet qui intéresse beaucoup de chercheurs et qui a fait l'objet de nombreux travaux. Plusieurs méthodes ont été proposées pour la détection d'anomalies et chaque méthode a ses forces et ses faiblesses. Patcha et Park (2007) ont fait une revue des méthodes utilisées pour la détection d'intrusion. Une revue plus générale des techniques existantes couvrant plusieurs approches est proposée dans Aggarwal (2017) et Chandola et al. (2009). Gupta et al. (2014) fait l'état de l'art des méthodes en fonction du type de données considérées : les données temporelles telles que les séries temporelles, les données spatio-temporelles et les flux de données. Salehi et Rashidi (2018), Souiden et al. (2016), Thakkar et al. (2016), Tellis et D'Souza (2018)

présentent également des méthodes applicables aux flux de données. Dans le Tableau 1, nous présentons une synthèse qui s'appuie sur 10 revues majeures dans la littérature. Nous identifions respectivement les techniques de détection d'anomalies, les types de jeux de données et les domaines d'application abordés dans chacune de ces 10 revues.

Le but de ce papier est de fournir un état de l'art complet en agrégeant plusieurs informations sur les différentes méthodes de détection d'anomalies, les jeux de données et les domaines d'applications. Une classification est proposée afin de recommander des méthodes de détection d'anomalies à utiliser selon le type de données dont on dispose (flux de données, série temporelle, graphes…) avec des références bibliographiques pertinentes (Tableau 2) et selon l'approche qu'on voudrait utiliser (Figure 1). Une exposition des forces et faiblesses de différentes techniques est disponible dans Aggarwal (2017), Chandola et al. (2009), Chalapathy et Chawla (2019).

		1	2	3	4	5	6	7	8	9	10
Techniques	Statistique	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	Clustering	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	Plus proches voisins	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	Classification	✓	✓	✓	✓		✓				
	Régression				✓					✓	
	Approche spectrale			✓	✓						
	Motifs fréquents				✓		✓				
	Deep learning				✓						✓
Type de jeux de données											
	Flux de données				✓	✓	✓	✓	✓	✓	
	Séries temporelles				✓	✓					✓
	Graphes					✓			✓		
	Grande dimension				✓	✓				✓	
	Séquentielles				✓						
	Spatio-temporelles				✓	✓					✓
	Spatiales				✓						
Domaines d'application											
	Détection d'intrusion		✓	✓	✓	✓					✓
	Détection de fraude			✓	✓						✓
	Santé			✓	✓	✓					✓
	Maintenance prédictive			✓	✓	✓					✓
	Réseaux de capteurs			✓	✓	✓				✓	✓
	Traitement d'images			✓	✓						✓
	Traitement de texte			✓	✓						
	Données biologiques					✓					
	Astronomie					✓					
	Économie					✓					

TAB. 1 – *Synthèse de 10 revues existantes : 1- Hodge et Austin (2004) 2- Patcha et Park (2007) 3- Chandola et al. (2009) 4- Aggarwal (2017) 5-Gupta et al. (2014) 6-Souiden et al. (2016) 7-Tellis et D'Souza (2018) 8-Salehi et Rashidi (2018) 9- Zhang (2013) 10-Chalapathy et Chawla (2019).*

3 Classification des méthodes

3.1 Les domaines d'application

La détection d'anomalies est transversale à tout domaine qui exploite les données. Ainsi, elle a de nombreuses applications possibles. Les domaines d'application ayant leur spécificité en fonction des données générées ou exploitées, toutes les méthodes de la détection d'anomalies ne sont pas adaptées à tous les domaines d'application. Chandola et al. (2009) et Aggarwal (2017) ont fait une revue couvrant plusieurs domaines d'application. Dans Gupta et al. (2014), les auteurs ont fait la revue des méthodes de détection d'anomalies temporelles applicables dans plusieurs domaines différents. La détection d'intrusion consiste à l'analyse d'une cible généralement un réseau ou un hôte pour détecter les comportements anormaux (Chandola et al. (2009)). Il s'agit en effet de tentatives frauduleuses d'accès à une ressource par violation de la sécurité mise en place pour la cible en question (Aggarwal (2017)). La détection de fraude permet de repérer les activités suspectes menées par un individu généralement sous une fausse identité (identité usurpée). Le Tableau 1 liste plusieurs revues ayant abordé différents domaines d'application tels que la détection d'intrusion, la détection de fraude, la santé, la maintenance prédictive dans l'industrie, les réseaux de capteurs, le traitement d'images, le traitement de texte, les données biologiques, l'astronomie et l'économie.

3.2 Les types de jeux de données

L'émergence des nouvelles technologies a conduit à la génération de différents types de données. On trouve plusieurs jeux de données ayant des caractéristiques différentes et apportant de nouveaux challenges dans la détection d'anomalies. Nous présentons dans le Tableau 1 les types de jeux de données abordés dans les 10 revues étudiées. Différents algorithmes de détection d'anomalies sont appliqués en fonction du type de jeux de données considéré. Nous listons dans le tableau 2 les algorithmes les plus utilisés pour chacun des types de jeux de données.

A la différence des séries temporelles, les flux de données sont générés de façon continue et infinie à une vitesse variable. Compte tenu de la taille des données produites, les flux de données ne peuvent pas être exhaustivement stockés pour une exploitation future. Les méthodes de détection d'anomalies dans les flux de données doivent donc être appliquées en temps réel souvent sans aucune connaissance a priori sur la distribution des données (Gupta et al. (2014) ; Tellis et D'Souza (2018) ; Salehi et Rashidi (2018)). Le traitement en ligne exige un algorithme de détection de faible complexité pour une exécution plus rapide que la vitesse d'arrivée des données et souvent aussi une faible consommation mémoire si la solution est implémentée sur un équipement à ressources limitées.

3.3 Les techniques de détection d'anomalies

Les techniques de détection d'anomalies existantes se basent sur deux propriétés importantes des anomalies : elles ont un comportement très différent des autres et elles sont rares. Parmi ces techniques, nous distinguons :

Jeux de données	Méthodes applicables	Références
Flux de données	SmartSifter, AnyOut, CluStream, LEAP, MiLOF, DCLUST, IncLOF, DenStream, Abstract-C, AMCOD,HPStream, WaveCluster, MDEF	Yamanishi et al. (2004) ; Aggarwal (2017) ; Cao et al. (2006) ; Chen et Tu (2007) ; Ren et Ma (2009) ; Gupta et al. (2014) ; Mishra et Chawla (2019) ; Zhang (2013)
Séries temporelles	ARMA, ARIMA, VARMA, CUSUM, EWMA, LSA, MLP, ART NN, AE,GAN	Gupta et al. (2014) ; Chalapathy et Chawla (2019)
Graphes	DBMM, ECOutlier, NetSpot, ParCube, Com2, NetSmile, DeltaCon	Salehi et al. (2014) ; Salehi et Rashidi (2018) ; Gupta et al. (2014, 2012)
Grande dimension	GLOF, HighDoD, SOF, CLIQUE, HPStream, ABOD, SOD, SPOT	Mishra et Chawla (2019) ; Domingues et al. (2018)
Séquentielles	CLUSEQ, TARZAN	Aggarwal (2017)
Spatio-temporelles	Outstretch, TRAOD, LSTM, CNN	Wu et al. (2008) ; Chalapathy et Chawla (2019)
Données spatiales	Moran scatterplot, DBSCAN	El Sibai et al. (2018) ; Ester et al. (1996)

TAB. 2 – *Méthodes applicables pour chaque type de jeu de données.*

— **Les techniques statistiques** qui peuvent être paramétriques ou non paramétriques. A la différence de l'approche non paramétrique, l'approche paramétrique suppose une connaissance a priori de la distribution des données. Les méthodes statistiques construisent un modèle avec un intervalle de confiance à partir des données existantes. Les nouvelles données qui ne correspondent pas à ce modèle seront considérées anormales (Desforges et al. (1998) ; Aggarwal (2017)).

— **Les techniques basées sur la proximité** qui regroupent celles basées sur les plus proches voisins et celles basées sur le clustering. Les techniques basées sur les plus proches voisins déterminent pour une observation o ses k plus proches voisins à travers le calcul de la distance entre toutes les observations du jeu de données. Ces méthodes nécessitent un calcul préalable, et de ce fait, elles sont coûteuses en temps d'exécution. Il existe deux approches de méthodes basées sur les plus proches voisins : l'approche basée sur la distance (Angiulli et Pizzuti (2002) ; Yamanishi et al. (2004)) et l'approche basée sur la densité (Breunig et al. (2000)). Les techniques de clustering ont pour objectif principal de diviser le jeu de données en clusters contenant les données qui ont des comportements similaires. On distingue deux approches dans ces techniques : l'approche basée sur la distance selon laquelle le cluster le plus éloigné représente une anomalie et l'approche basée sur la densité définie l'anomalie par le cluster qui contient le moins de données.

— **Les techniques basées sur le deep learning** qui représentent une classe d'algorithmes d'apprentissage automatique supervisé ou non supervisé basés sur

l'utilisation de plusieurs couches d'unité de traitement non linéaire. Parmi ces méthodes on cite les auto-encoders (AE) et One-Class Neural Networks (OC-NN) (Chalapathy et Chawla (2019)).

— **D'autres techniques** existent comme celles basées sur les machines à vecteurs de support (Schölkopf et al. (2000), les réseaux de neurones (Hodge et Austin (2004)), les méthodes adaptées aux grandes dimensions par construction de sous-espaces ou par réduction de dimension (Aggarwal (2017)).

Dans la Figure 1, nous présentons une synthèse de cette classification avec des exemples d'algorithmes appartenant à chacune de ces catégories.

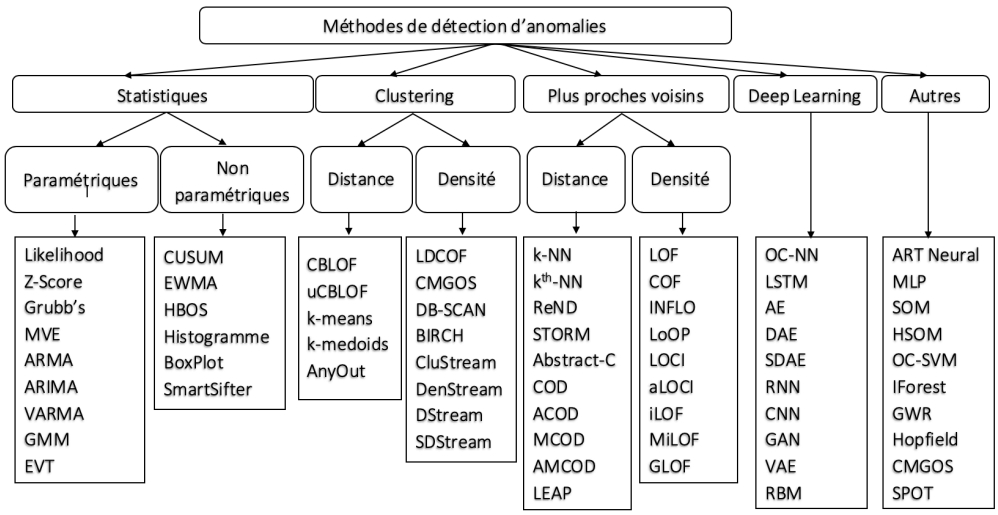

FIG. 1 – *Classification des différentes techniques de détection d'anomalies.*

4 Etude expérimentale

Dans cette partie, nous avons fait le choix de comparer expérimentalement trois méthodes de détection d'anomalies très utilisées dans la littérature : LOF, OC-SVM et Isolation Forest. En effet, il s'agit de méthodes performantes appartenant à des catégories différentes. Ces trois méthodes donnent de très bons résultats comparées aux autres méthodes de leurs approches respectives. De plus, elles sont souvent utilisées comme références de comparaison pour évaluer les performances des nouvelles méthodes de détection d'anomalies. Nous commençons par décrire chacune de ces méthodes puis nous présentons les jeux de données utilisés ainsi que les résultats obtenus.

4.1 Description des algorithmes : LOF, Isolation Forest et OC-SVM

Local Outlier Factor (LOF) est une méthode phare de la détection d'anomalies locales basée sur la densité de l'observation en question par rapport à la densité de ses plus proches voisins. Proposée par Breunig et al. (2000), LOF est une méthode non supervisée qui donne un score représentant le degré d'aberrance de l'observation. Les observations dont le degré d'aberrance est largement supérieur à 1 sont considérées comme anomalies. La méthode prend en paramètre le nombre de plus proches voisins à considérer k, et calcule le degré d'aberrance d'un point p par la formule suivante : $LOF_k(p) = \frac{\sum_{o \in N(p,k)} \frac{lrd_k(o)}{lrd_k(p)}}{k}$ avec $N(p,k)$ l'ensemble des k plus proches voisins de p. $lrd(p)$ est la densité d'accessibilité locale de p et correspond à l'inverse de la distance d'accessibilité moyenne entre p et ses k plus proches voisins : $lrd_k(p) = \frac{k}{\sum_{o \in N(p,k)} reach\_dist_k(p,o)}$ La distance d'accessibilité de p par rapport à o considérant les k plus proches voisins ($reach\_dist_k(p,o)$) entre deux points (p et o), correspond au maximum entre la distance entre le $k^{ième}$ plus proche voisin de o ($k-distance(o)$) et la distance entre p et o ($d(p,o)$), soit : $reach\_dist_k(p,o) = \max(k-distance(o), d(p,o))$. On trouve dans la littérature différentes améliorations de LOF comme Incremental LOF «iLOF»(Pokrajac et al. (2007)) qui est adaptée aux flux de données, mais qui consomme beaucoup de mémoire pour le calcul de la densité des nouvelles données entrantes. Memory Efficient ILOF «MILOF»(Salehi et al. (2016)) est une évolution de iLOF qui réduit la consommation mémoire tout en ayant une précision similaire à iLOF. Pour la détection d'anomalies dans les jeux de données de grandes dimensions, Lee et Cho (2016) ont proposé Grid-LOF «GLOF» qui divise le jeu de données en petites régions (Grid) avant de calculer la densité.

Isolation Forest ou IForest (Liu et al. (2008, 2012)) est une méthode basée sur les arbres de décision et les forêts aléatoires. Elle utilise l'isolation d'observations à partir de la construction de plusieurs arbres aléatoires. Quand une forêt d'arbres aléatoires et indépendants produit collectivement un chemin d'accès court pour atteindre une observation depuis la racine, celle-ci a une forte probabilité d'être une anomalie. Le nombre d'arbres utilisés est donc un important paramètre pour IForest. Le seuil de la détection est aussi un paramètre clé, il est donné par le score calculé pour chaque observation relativement aux autres observations. Si ce score est proche de 1 alors l'observation est considérée comme anomalie. Considérons un jeu de données de n observations et une observation x, le score $s(x,n)$ d'aberrance de x est calculé par la formule suivante : $s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$ avec $h(x)$ la longueur du chemin entre la racine de l'arbre et x. $E(h(x))$ est la moyenne des $h(x)$ de toute la forêt d'arbres. $c(n)$ représente la longueur du chemin moyen d'une observation depuis la racine dans le jeu de données de n observations. Isolation Forest fait partie des méthodes de détection d'anomalies les plus récentes et les plus utilisées. Elle donne également de bons résultats pour les jeux de données de grandes dimensions.

One-Class Support Vector Machine (OC-SVM) est une méthode de détection d'anomalies qui applique des algorithmes de SVM au problème de One class classifi-

cation (OCC) proposée par Schölkopf et al. (2000, 2001). Le séparateur à vaste marge (SVM) appelé aussi machine à vecteurs de support est très utilisé pour l'apprentissage automatique du fait de sa puissance et de sa polyvalence (classification linéaire, non-linéaire, régression). OCC est une approche de classification semi-supervisée qui consiste à repérer toutes les observations appartenant à une classe précise connue pendant l'apprentissage, dans tout le jeu de données. L'idée clé de cette méthode est de trouver un hyperplan dans un espace de grande dimension qui sépare les anomalies des données normales.

4.2 Jeux de données

Afin d'évaluer les performances des méthodes choisies en fonction de la dimension du jeu de données, nous avons utilisé deux jeux de données (KDD-Cup99 et Shuttle) très exploités par la communauté de détection d'anomalies pour l'étude comparative des méthodes. KDD-Cup99 HTTP a été conçu et publié par Goldstein et Uchida (2016) après quelques manipulations sur le jeu de données original KDD-Cup99 (Lazarevic et al. (2003)). Nos travaux ont porté sur une extraction de ce jeu de données. Statlog Shuttle a été obtenu par Goldstein et Uchida (2016) à partir du jeu de données original (Martin et al. (2007)) après réduction du nombre d'anomalies. Ces jeux de données rendus disponibles [1] contiennent un attribut qui indique la classe de chaque observation (anomalie ou non). Ils peuvent être utilisés pour les méthodes supervisées, semi-supervisées et non supervisées de détection d'anomalies. Le Tableau 3 regroupe les détails sur ces deux jeux de données de tailles et de dimensions différentes. Les différentes caractéristiques de ces deux jeux de données permettent d'évaluer l'impact de l'augmentation du volume et des dimensions des données sur la performance de la détection des différents algorithmes ainsi que sur leurs temps d'exécution.

Jeu de données	Taille(observations)	Attributs	Anomalies(observations)
Statlog Shuttle	46 464	10	878
KDD-Cup99 HTTP	103 351	30	176

TAB. 3 – *Caractéristiques des deux jeux de données utilisés.*

4.3 Résultats

Plusieurs critères peuvent être considérés pour comparer les performances des méthodes de détection d'anomalies. Pour qualifier la précision de la détection, nous nous intéressons aux 3 mesures suivantes :

— L'aire sous la courbe ROC (**ROC AUC**) qui est un standard dans la comparaison des performances des méthodes de détection d'anomalies,

— **La Spécificité** : $Spécificité = \frac{VN}{VN + FP}$

— **Le Rappel** : $Rappel = \frac{VP}{VP + FN}$

1. http://dx.doi.org/10.7910/DVN/OPQMVF

Dans ce contexte, les positifs représentent les données anormales et les négatifs représentent les données normales. Nous avons également mesuré le **temps d'exécution** qui dépend à la fois de la complexité de l'algorithme et de la taille du jeu de données considéré.

Le tableau 4 résume les résultats des 3 méthodes sur les deux jeux de données considérés.

Jeu de données	Rappel	Spécificité	ROC AUC	CPU Time(s)
IForest(SSh)	**0.98**	**0.92**	**0.95**	**3.86**
IForest(KDD)	**0.90**	**1**	**0.95**	**28.92**
LOF(SSh)	0.38	0.91	0.65	32.18
LOF(KDD)	**0.90**	0,67	0.79	35.98
OC-SVM(SSh)	0.95	0.60	0.78	253.67
OC-SVM(KDD)	0.58	0.51	0.54	4615.74

TAB. 4 – *Tableau comparatif de la performance de IForest, LOF et OC-SVM sur les jeux de données Shuttle(SS) et KDD-Cup99 HTTP(KDD).*

IForest a une durée d'exécution plus petite que LOF et OC-SVM pour les deux jeux de données de tailles et de dimensions différentes. Cela s'explique par le fait que IForest ne calcule ni la distance entre les points ni la densité des points. IForest détecte un grand nombre d'anomalies avec beaucoup moins de fausses alertes que OC-SVM qui dans cet exercice est dépassé par LOF. LOF laisse passer beaucoup d'anomalies (plus de 60% pour shuttle) mais fait beaucoup moins de fausses alertes que OC-SVM qui par contre détecte 95% des anomalies de Shuttle. Il faut noter également qu'en terme de temps d'exécution, LOF est beaucoup plus rapide que OC-SVM. En général, les performances de toutes ces méthodes régressent (mais pas de la même manière) au fur et à mesure que la taille du jeu de données et ses dimensions augmentent. OC-SVM est adaptée aux jeux de données complexes de grandes dimensions, mais n'est pas adaptée aux jeux de données de grande taille. LOF est adaptée à la détection d'anomalies locales, car l'observation n'est comparée qu'à ses k plus proches voisins sans avoir de vision globale sur la totalité des données. Le mode de fonctionnement de Isolation Forest (pas de calcul de distance, pas de calcul de densité ni recherche de modèle) fait que le traitement est rapide même pour un jeu de données de grande taille ou de grande dimension. La contribution de différents arbres indépendants dans la prise de décision fournit également une très bonne précision dans la détection.

5 Conclusion

La détection d'anomalies étant transversale à tout domaine de traitement de données, différentes méthodes sont proposées suivant les contraintes de chaque domaine d'application et type de données. Dans ce papier, nous avons fait une revue des méthodes existantes et celles adaptées à chaque domaine d'application et principaux types de jeux de données. Nous avons fourni une classification de ces méthodes suivant différentes approches. L'étude comparative que nous avons menée a montré que IForest

est plus performant que LOF et OC-SVM en termes de précision de la détection et de durée d'exécution sur deux jeux de données de différentes tailles et dimensions.

La détection d'anomalies dans les flux de données impose des contraintes fortes notamment un traitement en temps réel souvent sans aucune connaissance préalable sur les données. On a également besoin de concevoir des algorithmes de détection distribués pour faire face au grand débit, toujours croissant, des flux de données. Nos futurs travaux de recherche s'intéresseront à ces problématiques qui ne sont pas suffisamment explorées aujourd'hui.

Références

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

Angiulli, F. et C. Pizzuti (2002). Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 15–27. Springer.

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

Cao, F., M. Estert, W. Qian, et A. Zhou (2006). Density-based clustering over an evolving data stream with noise. In *Proceedings of the 2006 SIAM international conference on data mining*, pp. 328–339. SIAM.

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

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

Chen, Y. et L. Tu (2007). Density-based clustering for real-time stream data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142. ACM.

Desforges, M., P. Jacob, et J. Cooper (1998). Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C : Journal of Mechanical Engineering Science 212*(8), 687–703.

Domingues, R., M. Filippone, P. Michiardi, et J. Zouaoui (2018). A comparative evaluation of outlier detection algorithms : Experiments and analyses. *Pattern Recognition 74*, 406–421.

Dunning, T. et E. Friedman (2014). *Practical machine learning : a new look at anomaly detection.* " O'Reilly Media, Inc.".

El Sibai, R., Y. Chabchoub, et C. Fricker (2018). Using spatial outliers detection to assess balancing mechanisms in bike sharing systems. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pp. 988–995. IEEE.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Volume 96, pp.

226–231.

Goldstein, M. et S. Uchida (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one 11*(4), e0152173.

Gupta, M., J. Gao, C. C. Aggarwal, et J. Han (2014). Outlier detection for temporal data : A survey. *IEEE Transactions on Knowledge and Data Engineering 26*(9), 2250–2267.

Gupta, M., J. Gao, Y. Sun, et J. Han (2012). Integrating community matching and outlier detection for mining evolutionary community outliers. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 859–867. ACM.

Hawkins, D. M. (1980). *Identification of outliers*, Volume 11. Springer.

Hodge, V. et J. Austin (2004). A survey of outlier detection methodologies. *Artificial intelligence review 22*(2), 85–126.

Lazarevic, A., L. Ertoz, V. Kumar, A. Ozgur, et J. Srivastava (2003). A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pp. 25–36. SIAM.

Lee, J. et N.-W. Cho (2016). Fast outlier detection using a grid-based algorithm. *PloS one 11*(11), e0165972.

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.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD) 6*(1), 3.

Martin, R., M. Schwabacher, N. Oza, et A. Srivastava (2007). Comparison of unsupervised anomaly detection methods for systems health management using space shuttle. In *Main Engine Data," Proceedings of the Joint Army Navy NASA Air Force Conference on Propulsion, 2007*. Citeseer.

Mishra, S. et M. Chawla (2019). A comparative study of local outlier factor algorithms for outliers detection in data streams. In *Emerging Technologies in Data Mining and Information Security*, pp. 347–356. Springer.

Patcha, A. et J.-M. Park (2007). An overview of anomaly detection techniques : Existing solutions and latest technological trends. *Computer networks 51*(12), 3448–3470.

Pokrajac, D., A. Lazarevic, et L. J. Latecki (2007). Incremental local outlier detection for data streams. In *2007 IEEE symposium on computational intelligence and data mining*, pp. 504–515. IEEE.

Ren, J. et R. Ma (2009). Density-based data streams clustering over sliding windows. In *2009 Sixth international conference on fuzzy systems and knowledge discovery*, Volume 5, pp. 248–252. IEEE.

Salehi, M., C. Leckie, J. C. Bezdek, T. Vaithianathan, et X. Zhang (2016). Fast memory efficient local outlier detection in data streams. *IEEE Transactions on Knowledge and Data Engineering 28*(12), 3246–3260.

Salehi, M., C. A. Leckie, M. Moshtaghi, et T. Vaithianathan (2014). A relevance

weighted ensemble model for anomaly detection in switching data streams. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 461–473. Springer.

Salehi, M. et L. Rashidi (2018). A survey on anomaly detection in evolving data :[with application to forest fire risk prediction]. *ACM SIGKDD Explorations Newsletter 20*(1), 13–23.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, et R. C. Williamson (2001). Estimating the support of a high-dimensional distribution. *Neural computation 13*(7), 1443–1471.

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.

Souiden, I., Z. Brahmi, et H. Toumi (2016). A survey on outlier detection in the context of stream mining : review of existing approaches and recommadations. In *International Conference on Intelligent Systems Design and Applications*, pp. 372–383. Springer.

Tellis, V. M. et D. J. D'Souza (2018). Detecting anomalies in data stream using efficient techniques : A review. In *2018 International Conference on Control, Power, Communication and Computing Technologies (ICCPCCT)*, pp. 296–298. IEEE.

Thakkar, P., J. Vala, et V. Prajapati (2016). Survey on outlier detection in data stream. *Int. J. Comput. Appl 136*, 13–16.

Wu, E., W. Liu, et S. Chawla (2008). Spatio-temporal outlier detection in precipitation data. In *International Workshop on Knowledge Discovery from Sensor Data*, pp. 115–133. Springer.

Yamanishi, K., J.-I. Takeuchi, G. Williams, et P. Milne (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery 8*(3), 275–300.

Zhang, J. (2013). Advancements of outlier detection : A survey. *ICST Transactions on Scalable Information Systems 13*(1), 1–26.

Summary

Anomaly detection is an important issue in many application domains. For example, cybercrime can cause considerable economic losses and threaten companies survival. Securing its information system has become a priority and a strategic issue for all types of companies. Other areas are also be impacted such as health, transport, etc. The implemented supervision solutions are often based on anomaly detection algorithms from datamining and machine learning domains. We present in this paper a complete state of the art on anomaly detection algorithms. We propose a classification of these methods based on the type of data sets (flows, time series, graphs, etc.), the application domain and the considered approach (statistics, classification, clustering, etc.). We then focus on three algorithms: LOF, OC-SVM and Isolation Forest, that we test on two different datasets to compare their performance.