Évaluation comparative de méthodes non supervisées pour la détection de points anormaux dans les flux de données

Anne Marthe Sophie Ngo Bibinbe, Michael Franklin Mbouopda,
Gertrude Raissa Mbiadou Saleu, Engelbert Mephu Nguifo

Université Clermont Auvergne, Clermont Auvergne INP, CNRS,
Mines Saint-Etienne, LIMOS, 63000 Clermont-Ferrand, France
{anne.ngo_bibinbe, engelbert.mephu_nguifo}@uca.fr

1 Introduction

Il existe plusieurs méthodes basées sur des hypothèses variées pour la détection d'anomalies dans les flux de données. Le choix d'une méthode est lié à ses performances sur des types de données spécifiques. Les flux de données peuvent être caractérisés par la présence de saisonalité, tendance, cycle et Concept drift (changement des propriétés statistiques des données). Dans ce travail, nous comparons suivant la latence (temps de traitement d'une instance) et performance un ensemble de méthodes de détection d'anomalies dans les flux de données avec des hypothèses diverses sur des jeux de données aussi bien univariés que multivariés (sur lesquels nous avons identifié les caractéristiques présentes).

2 Experimentation et discussion

Nous avons choisi les méthodes suivantes pour leur efficacité reportée dans les travaux de l'état de l'art : MILOF, Online ARIMA, KitNet, IforestASD et HStree. Les jeux de données dont les causes d'anomalies sont connues issus des banques de données SKAB (multivarié) et NAB (univarié) ont été utilisés. Des explications et résultats plus détaillés des méthodes ainsi que leurs références et celles des jeux de données peuvent être trouvées sur notre github (Ngo Bibinbe et al., 2021).

La métrique d'évaluation utilisée pour comparer les méthodes est le F1-score. Une méthode trouve une anomalie si elle en détecte à 1% (de la longueur de la série) près de la position définie de l'anomalie (Nakamura et al., 2020). Pour chaque jeu de données, les meilleurs hyperparamètres des méthodes sont trouvés par optimisation bayésienne. KitNet a été testé en multivarié et Online ARIMA en univarié ceci dû à leurs contraintes de conception.

Le tableau 1 résume le nombre de jeux de données où les méthodes ont eu les meilleurs scores, et parmi ceux-là le nombre ayant des concept drift, saisonnalités, tendances et cycles (sachant qu'un jeu de données peut avoir plus d'une seule des caractéristiques possibles). Le tableau 2 résume les temps moyens de réponse des méthodes pour chaque nouvelle instance.

Ce travail a été soutenu par le PIA LabEx IMobS3, et le projet DASMA financé par la BPI/Pfeiffer Vacuum. Nous remercions également les relecteurs anonymes pour leurs remarques constructives.

Méthode	Meilleur score	Concept drift	Saisonalité	Tendance	Cycle
MILOF	1/14	0/1	0/1	1/1	1/1
HS-tree	7/14	5/7	4/7	5/7	2/7
iForestASD	3/14	2/3	1/3	2/3	0/3
Online ARIMA	3/7	2/3	3/3	3/3	2/7
KitNet	2/7	0/2	2/2	0/2	0/2

TAB. 1 – *Récapitulatif des observations (scores, caractéristiques)*

	MILOF	iForestASD	HS-tree	Online ARIMA	KitNet
univariées (ms)	22.2	27.8	222.8	11.06	-
multivariées (ms)	9.5	31.9	80.7	-	0.32

TAB. 2 – *Temps moyen de réponse (latence) des méthodes sur les jeux de données*

Nous remarquons qu'Online ARIMA donne les meilleurs scores en univarié en présence de tendances et saisonnalités. Ceci pourrait être dû au fait qu'il arrive à capturer les dépendances temporelles par la prévision. De façon générale, les méthodes basées sur les arbres donnent de bons scores en univarié et multivarié. Nous remarquons que les méthodes faisant de la descente de gradient (KitNet et Online ARIMA) ont des temps de latence plus petits que les autres méthodes. Dans les résultats détaillés des méthodes (Ngo Bibinbe et al., 2021), nous remarquons que MILOF a des plus faibles scores en multivarié.

3 Conclusion

Cette expérimentation met en exergue : l'efficacité de la descente de gradient dans le temps de latence des méthodes, qu'Online ARIMA par son approche prévisionnelle, capture la saisonnalité et la tendance ; et les bons scores des méthodes basées sur les arbres de façon générale en présence de chacune des caractéristiques.

Références

Nakamura, T., M. Imamura, R. Mercer, et E. Keogh (2020). Merlin : Parameter-free discovery of arbitrary length anomalies in time series archives. In *IEEE (ICDM)*, pp. 1190–1195.

Ngo Bibinbe, A. M. S., M. F. Mbouopda, G. R. Mbiadou Saleu, et E. Mephu Nguifo (2021). url https ://github.com/nams2000/anomaly-detection-in-data-stream.

Summary

In this paper, we compare some unsupervised data stream abnormal point detection methods with emphasis on their performance and response time, as well as the presence of concept drift, seasonality, trend, cycle as characteristics of tested datasets.