myCADI: my Contextual Anomaly Detection using Isolation

Véronne Yepmo, Grégory Smits

▶ **To cite this version:**

Véronne Yepmo, Grégory Smits. myCADI: my Contextual Anomaly Detection using Isolation. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), Oct 2024, Boise, ID, United States. 10.1145/3627673.3679208 . hal-04743207

HAL Id: hal-04743207
https://hal.science/hal-04743207v1

Submitted on 18 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

myCADI: my Contextual Anomaly Detection using Isolation

Véronne Yepmo
veronne.yepmo-tchaghe@irisa.fr
Université de Rennes - IRISA
Lannion, France

Grégory Smits
gregory.smits@imt-atlantique.fr
IMT Atlantique - Lab STICC
Brest, France

ABSTRACT

myCADI is a machine learning framework associated with a graphical interface for discovering and understanding the internal structure of an unsupervised dataset. It is an intuitive end-user interface to the CADI approach [9], which uses a revised version of the Isolation Forest (IF) method to both 1) identify local anomalies, 2) reconstruct the cluster-based internal structure of the data, and 3) provide end-users with explanations of how anomalies deviate from the found clusters. myCADI takes numerical data as input and is structured around several interfaces, each of which displays a ranked list of the found anomalies, a description of the subspaces in which the different clusters lie, and feature attribution explanations to ease the interpretation of anomalies. These explanations make explicit why a selected point is considered to be a local anomaly of one (or more) cluster(s). The framework also provides dataset and trees visualizations.

CCS CONCEPTS

• **Computing methodologies** → **Anomaly detection**; **Cluster analysis**; *Artificial intelligence*; • **Information systems** → **Data analytics**.

KEYWORDS

Anomaly detection, Robust clustering, Anomaly explanation, XAI

ACM Reference Format:
Véronne Yepmo and Grégory Smits. 2024. myCADI: my Contextual Anomaly Detection using Isolation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679208

1 INTRODUCTION

Faced with a raw dataset to analyze, users need tools to help them quickly understand the main trends in the data before deciding to invest time and energy in deeper and more specific treatments. Different types of knowledge are particularly useful in providing users with a complete insight into the data: Are there points that are so different that they can be considered anomalies? What is the internal structure of the regular points? What are the relationships between anomalies and regularities? Many Machine Learning (ML) techniques: anomaly detectors, clustering algorithms and post-hoc explainers, have been designed to answer these three questions independently. Furthermore, anomaly detection and clustering are usually considered as disjoint problems. Anomalies, i.e. deviating instances, are therefore often discarded during clustering, or anomaly detection solely is performed, resulting in a binary partition of the dataset (anomalies vs regular instances). However, some instances deviate only from a portion of the dataset. These instances, called local anomalies are acknowledged in the anomaly detection literature, with dedicated methods like LOF [2]. However, as a result of the handling of anomaly detection and clustering as opposite problems, the local anomalies are explained as if they were deviating from all the regular instances. Post-hoc anomaly explanation methods like COIN [6] and ATON [8] tackle this issue by performing (COIN) a clustering of the regular neighbors of the anomaly while generating the explanation. However, these two approaches do not provide any information regarding the eventual clusters located in the dataset, and rely on external clustering algorithms. To have a complete view of the data, the user therefore needs to tune separately three different algorithms for each task, and then to combine them in a pipeline.

The proposed demo introduces a web application, called myCADI that provides all these data analysis functionalities based on a same inferred data model. myCADI relies on the CADI ML framework, which stands for Contextual Anomaly Detection using an Isolation forest. Based on a revisited version of the initial Isolation Forest (IF) method introduced in [5], myCADI still extracts anomalies efficiently, while also reconstructing a cluster-based structure of the regular points, these two data analysis tasks being performed by leveraging the inferred IF only. In addition, the IF is also used to exhibit links between anomalies and the found clusters in order to discriminate between global and local anomalies.

As a complement to the performances of myCADI already commented in [9], the demonstration of CADI aims at illustrating the usefulness and interpretability of the extracted knowledge. A widely accessible data analysis scenario on the statistics of basketball players is used during the demonstration showing that the Most Valuable Players (MVP) are anomalies and you will understand why.

2 CADI

This section first briefly presents the ML approach to local anomaly detection and explanation. The described functionalities are provided as a Python API publicly available on GitLab [1].

2.1 Data Process Pipeline

CADI helps end-users discriminate between global and local anomalies. Whereas the severity of a global anomaly is calculated based on a comparison wrt. the rest of the dataset, a local anomaly is characterized by the extend to which it deviates from its close

This work is licensed under a Creative Commons Attribution-ShareAlike International 4.0 License.

CIKM '24, October 21–25, 2024, Boise, ID, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10.
https://doi.org/10.1145/3627673.3679208

[1] https://gitlab.com/yveronne/cadi

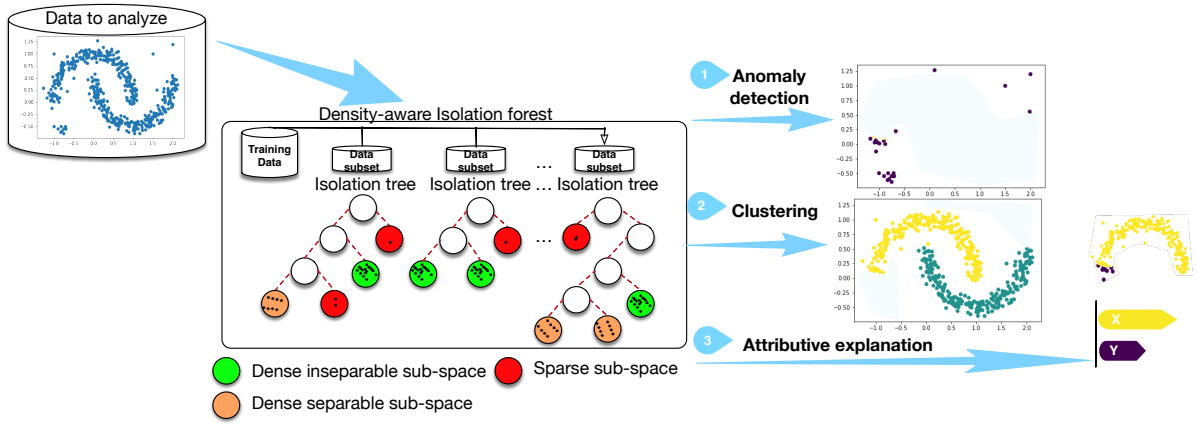


Figure 1: Leveraging a density-aware IF to anomaly detection, clustering and anomaly explanation

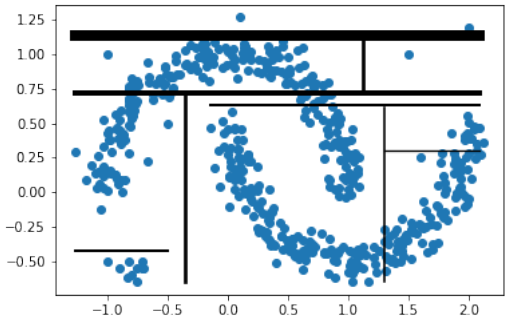


Figure 2: Density-aware isolation tree on a toy 2D dataset

neighborhood. To provide users with a complete overview of the dataset they have to analyze, found anomalies are compared with the cluster-based inner structure of the data. Instead of combining different ML techniques to first extract the anomalies, then to partition the data and finally to apply post-hoc explanation mechanisms, CADI relies on a same unified model inferred from the data to perform these three complementary tasks. The hypothesis at the origin of this approach, that we propose to illustrate through the proof-of-concept myCADI (Section 3), is that relying on a same data principle, namely points separability, eases the understanding of the different types of extracted knowledge: anomalies, clusters, their commonalities and differences.

As illustrated in Figure 1, a density-aware IF (Section 2.2) is first inferred from the training data. From this enhanced IF, anomalies are extracted and a cluster-based partition of the remaining points is reconstructed. Commonalities and differences between anomalies and clusters are also determined using the IF only and presented as attributive explanations.

2.2 Density-Aware IF

The corner stone of the proposed tool is an inferred data model composed of isolation trees, a data structure initially conceived to efficiently exhibit outliers from an unsupervised dataset [5]. An important pro of the IF approach to outlier detection is to not depend on a preselected distance measure as it relies on randomly generated splits of the universe, a split being defined as a couple (A, v) where A is an attribute and v a value in its domain. In order to be able to reconstruct the data inner structure in addition to isolate outliers, myCADI introduces a density criteria to discard splits that fall in dense regions of points and that may separate clusters to reconstruct. The impact of this density-based criteria is graphically illustrated in Figure 2 where kept separations surround dense areas. It has been shown in [9] that this criteria yields a clear distinction between leaves of the isolation trees containing points isolated in sparse regions from those containing inseparable groups of points located in dense regions.

2.3 Detecting and Explaining Local Anomalies

In the seminal IF approach, the anomaly score computed for each analyzed point depends on its average isolation depth in the different trees. The density constraint imposed on the randomly generated splits in CADI may lead to the isolation in a top leaf of a dense group of inseparable points, hence the use of a dedicated anomaly scoring strategy that is a function of the leaf cardinality instead of its depth. Consequently, the first feedback given to the end-user about the data is an ordered list of found anomalies. Leaves of the isolation trees containing inseparable groups of points correspond to (parts of) clusters. Using a strategy inspired from a grid-based clustering method [1], the inner structure of the dataset is reconstructed. A cluster is a connected component of a graph, see Figure 3, where nodes are dense leaves of the isolation trees connected by weighted links when their delimitation subspaces somewhat intersect, these delimitations being reconstructed from the IF.

As a unified data structure to answer different complementary data analysis needs, the density-aware IF is again exploited to provide end-users with contextual explanations [6] about each local anomaly. Thus, each anomaly is associated with a quantification

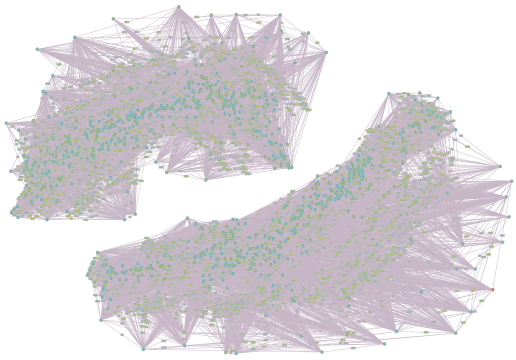


Figure 3: Graph of intersecting dense leaves

of the point abnormality, a local positioning wrt. its surrounding regularities (i.e. clusters), and attributive explanations. These explanations indicate the extent to which an attribute contributes to the abnormality of the point and to its local assignment(s) to found clusters. As illustrated in Figure 4, to determine the attributes that make of a point a local anomaly of a cluster, one leverages the paths in the isolation trees that go to the deepest common ancestor (blue hexagon in Fig. 4) of the leaves isolating the anomaly (red x_2 in Fig. 4) and the dense leaves of the cluster (green squared l_1 leaf in Fig. 4).

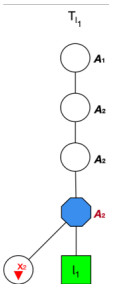


Figure 4: Attributive explanation generated from isolation paths

3 MYCADI

myCADI is a web-based graphical interface to CADI that gives an intuitive access to complementary data analysis functionalities: anomaly detection, clustering and anomaly explanation. In the existing literature, anomaly detection prototypes like the library PyOD [10] are limited to the identification of anomalies. Exathlon [3] and SEDAF [4] are prototypes including anomaly explanation. However, they are designed for time series for the former, and streaming data for the latter. Clustering explanation prototypes like Cluster-Explorer [7] on the other hand provide explanations to clusters. In contrast, myCADI showcases the importance of realizing the three tasks on tabular data, using the same model to have a better understanding of the dataset.

The graphical interface is structured in different panels dedicated to the dataset management, the IF, the anomalies, and the clusters.

3.1 Graphical Functionalities

myCADI is a React JS web application to manipulate in an intuitive way the ML functionalities provided by CADI. It allows users to 1) load a dataset in a csv format, 2) train a density-aware isolation forest, 3) extract anomalies, 4) reconstruct the inner structure of the data regularities, and 5) understand each found anomaly thanks to their associated local explanations. The goal of the demonstration is to convince data scientists and potential end-users of the interest of using a unified data model (i.e. a density aware isolation forest) and a unified data principle (i.e. points separability) to perform complementary data analysis tasks. The underlying hypotheses of the CADI approach is that the use of a unified data model to extract different types of knowledge makes their understanding easier. This is why for a selected outlier or cluster, from their respective visualization panels, one may then move to the IF panel to understand how their sparse or conversely dense subspaces have been identified thanks to recursive separations of the universe. Figure 5 shows these three main panels that provide functionalities to peruse the trees of the forest (top side of Figure 5), to go through the most suspicious anomalies (center of Figure 5) and to understand points distributions within each cluster (bottom side of Figure 5 with a focus on the *nbPoints* attribute of the dataset).

Figure 6 shows how local explanations make it possible to easily understand why a suspicious point is considered as a local anomaly of a particular cluster.

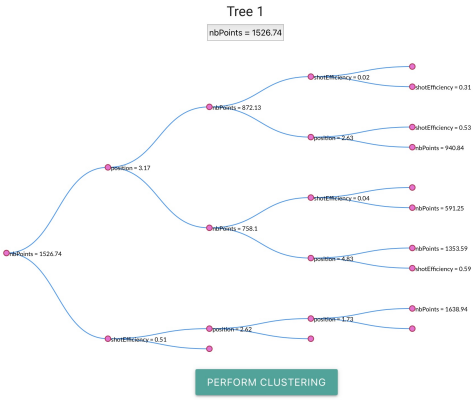
3.2 Demonstration Scenario

During the demo session, two use cases will be suggested. The first one aims at showing the efficiency and precision of the CADI approach with a comparison of several anomaly detectors (IF, LOF, and auto-encoders) and approaches (COIN [6] and ATON [8]) providing contextual explanations as well. Performance metrics (precision, recall and ARI) will allow users to check that CADI generally outperforms other approaches dedicated to anomaly detection or to post-hoc explanations. This "live" comparison with state-of-the-art approaches will be done on 13 datasets classically used among the anomaly detection community (*annthyroid, arrhythmia, breast, cover, hbk, http, ...*), as well as artificial data involving clusters and local anomalies to evaluate the explanation mechanism.

The content of these datasets is however abstruse for most of us. This is why a second demonstration scenario focuses on the analysis of a more accessible dataset, the statistics of each NBA player for the season 2019-2020 [2]. myCADI is applied on the attributes position (from 1-Point Guard to 5-Center), number of points and field goal percentage (renamed ShotEfficiency).

3.2.1 Building the forest. After selecting and loading the dataset, a forest needs to be built. The values of the hyper-parameters are specified. They are left to the default values, except the depth limit of each tree which is set to the maximum value, allowing the construction of fully grown trees. When the forest is built, the user can explore the different trees and see the separations selected during each step (Fig. 5 top).

[2] https://www.basketball-reference.com/leagues/NBA_2020_totals.html



Anomaly Scores

Instance #	Coordinates	Anomaly Score
12	[4.000e+00 5.530e-01 1.857e+03]	0.999844
258	[2.000e+00 4.440e-01 2.335e+03]	0.999648
68	[2.000e+00 4.890e-01 1.863e+03]	0.999258
169	[3.000e+00 5.310e-01 1.504e+03]	0.999258
40	[2.000e+00 4.550e-01 1.741e+03]	0.99918
335	[5.000e+00 5.280e-01 1.456e+03]	0.99918

Distribution of cluster 1

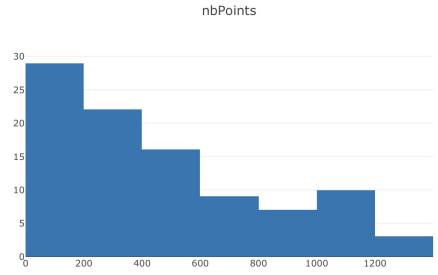


Figure 5: The IF (top), anomalies (center) and clusters (bottoms) panels

3.2.2 Computing anomaly scores. The user can ask for the computation of the anomaly scores to obtain a list of suspicious points from the most abnormal to the least abnormal point (Fig. 5 middle).

3.2.3 Clustering. In order to perform clustering, an anomaly score threshold needs to be selected, since clustering is performed on regular instances only. With the score selected, the method provides 7 clusters. The last two clusters contain respectively 1 and 2 points. The five main clusters obtained gather players having the same position, which is the most intuitive clustering of this dataset. This result shows that myCADI is able to extract meaningful clusters.

Explanation for instance 12

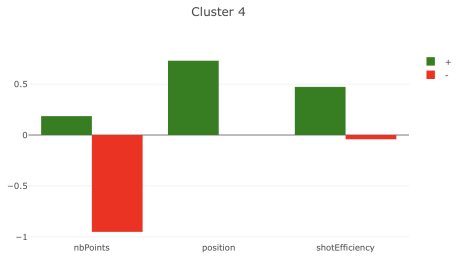


Figure 6: Local explanations confronting an outlier with its closest cluster

3.2.4 Explaining anomalies in relation to clusters. Figure 6 explains why the player 12 is considered as an anomaly. This player is Giannis Antetokounmpo, who plays Power Forward (PF, Position 4). According to myCADI, he is deviating from cluster 4, gathering Power Forwards. Giannis scored 1857 points during the season at 0.553 (efficiency). However, as shown by the data distribution of cluster 4, all regular PFs scored 1199 points or less, and most of them had an efficiency around 0.45 while doing so. Giannis scored much more points than the other PFs and was above the average in terms of efficiency, as well explained by Fig. 6. The major reason why Giannis is an anomaly in this dataset is not only because of the number of points. In fact, the scoring leader of the season was James Harden (2335 points). James Harden is a Shooting Guard (SG, Position 2). Even if 2335 is an exceptionally high number of points during a season (he is the second most abnormal instance), Shooting Guards are expected to score more points than players playing at other positions. What was really particular about Giannis during the 2019-2020 season was that he scored many points for a PF. He was named MVP that year. James Harden came third. So, if you want to know who will be elected as next season's MVP, just wait until the end of the regular season and ask myCADI.

4 CONCLUSION

This paper introduced myCADI, a web application allowing the users, given an unsupervised dataset, to identify deviating instances, cluster the regular instances and explain the abnormal instances in relation to the found clusters. Because the underlying principle solving these three tasks is the same, users are spared the choice of task-specific algorithms and the subsequent hyper-parameters tuning. Through a demonstration on real-world data, it was shown that myCADI is indeed able to extract intuitive groups and provide useful explanations regarding the abnormality of instances, therefore allowing a complete analysis of a dataset when the user is particularly interested about local anomalies.

ACKNOWLEDGMENTS

This research is part of the SEA DEFENDER project funded by the French DGA (Directorate General of Armaments).

REFERENCES

[1] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. 1998. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*. Association for Computing Machinery, New York, NY, USA, 94–105. https://doi.org/10.1145/276304.276314

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. Association for Computing Machinery, 93–104. https://doi.org/10.1145/342009.335388

[3] Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. 2021. A demonstration of the exathlon benchmarking platform for explainable anomaly detection. *Proceedings of the VLDB Endowment (PVLDB)* (2021).

[4] F Jiechieu Kameni, AMS Ngo Bibinbe, V Cako, AJ Djiberou Mahamadou, MR Bakari, KD Nguetche, D Kamga Nguifo, A Bertrand, MF Mbouopda, and R El Cheikh. 2024. SEDAF: Prototype d'un Système Explicable de Détection d'Anomalies dans les Flux de Données. (2024), 441–448.

[5] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data* 6 (mar 2012). https://doi.org/10.1145/2133360.2133363

[6] Ninghao Liu, Donghwa Shin, and Xia Hu. 2018. Contextual outlier interpretation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2461–2467. https://doi.org/10.5555/3304889.3305002

[7] Sariel Tutay and Amit Somech. 2023. Cluster-Explorer: An interactive Framework for Explaining Black-Box Clustering Results. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 5106–5110. https://doi.org/10.1145/3583780.3614734

[8] Hongzuo Xu, Yijie Wang, Songlei Jian, Zhenyu Huang, Yongjun Wang, Ning Liu, and Fei Li. 2021. Beyond Outlier Detection: Outlier Interpretation by Attention-Guided Triplet Deviation Network. In *Proceedings of the Web Conference 2021*. Association for Computing Machinery, New York, NY, USA, 1328–1339. https://doi.org/10.1145/3442381.3449868

[9] Véronne Yepmo, Grégory Smits, Marie-Jeanne Lesot, and Olivier Pivert. 2024. CADI: Contextual Anomaly Detection using an Isolation Forest. In *The 39th ACM/SIGAPP Symposium On Applied Computing*. 935–944.

[10] Yue Zhao, Zain Nasrullah, and Zheng Li. 2019. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research* 20, 96 (2019), 1–7. http://jmlr.org/papers/v20/19-011.html