A Classification of Anomaly Explanation Methods

Véronne Yepmo, Grégory Smits, Olivier Pivert

▶ **To cite this version:**

Véronne Yepmo, Grégory Smits, Olivier Pivert. A Classification of Anomaly Explanation Methods. Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI), Sep 2021, (Online), France. hal-03337036

HAL Id: hal-03337036
https://hal.science/hal-03337036v1

Submitted on 7 Sep 2021

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.

A Classification of Anomaly Explanation Methods

Véronne Yepmo Tchaghe, Grégory Smits, and Olivier Pivert

Univ Rennes, IRISA - UMR 6704, F-22305 Lannion, France
{veronne.yepmo-tchaghe,gregory.smits,olivier.pivert}@irisa.fr

Abstract. The usage of algorithms in real-world situations is strongly desired. But, in order to achieve that, final users need to be reassured that they can trust the outputs of algorithms. Building this trust requires algorithms not only to produce accurate results, but also to explain why they got those results. From this last problematic a new field has emerged: eXplainable Artificial Intelligence (XAI). Deep learning has greatly benefited from that field, especially for classification tasks. The considerable amount of works and surveys devoted to deep explanation methods can attest that. Other machine learning tasks, like anomaly detection, have received less attention when it comes to explaining the algorithms outputs. In this paper, we focus on anomaly explanation. Our contribution is a categorization of anomaly explanation methods and an analysis of the different forms anomaly explanations may take.

Keywords: anomaly explanation · outlier interpretation · XAI.

1 Introduction

An outlier/anomaly/irregularity is *an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism* [8]. Finding those deviating observations constitutes anomaly detection. Anomaly detection has many applications, ranging from spam detection in mail servers to the identification of cancerous cells in MRI photographs. It can be dealt with a binary classification task in which there is high imbalance between the classes (because anomalies are few in comparison to regular data points). But this requires knowing in advance the types of anomalies that can be found (even in real-world situations), which is not a suitable assumption because new anomalies, different from the ones learned, can appear after the model was developed. A more realistic design of anomaly detection uses density: low density regions are more likely to contain outliers (e.g: Local Outlier Factor (LOF)). Another outlier identification setting is to build a model for the regular instances and consider as anomalies instances which do not fit the model (e.g: clustering, isolation forest, one-class Support Vector Machines (one-class SVMs)).

In regular classification, when explaining an instance, the interest is on what makes the instance similar to the other instances of the same class; which common properties are shared by the instances of the same class. In contrast, in

anomaly explanation, the knowledge sought is about how the anomaly differs from the other instances. That is why anomaly explanation cannot be entirely managed like classification explanation and deserves a particular attention. This paper provides a taxonomy of anomaly explanation methods. After a review of the existing works on the topic in section 2, the proposed categorization will be detailed in section 3. For each category, its advantages and limits, the purpose of the explanations generated, along with some examples will be given.

2 Existing Comparison Criteria of Anomaly Explanations

Although there is no work entirely dedicated to a review of the anomaly explanation field, existing approaches to anomaly explanation are sometimes compared according to their properties and to the nature of the generated output. In [4] and [23], two types of anomaly explanation methods are mentioned: model-agnostic explanations and model-specific explanations. A model-specific method is a method developed for a particular machine learning algorithm, while a model-agnostic method can be used with any algorithm. In [4], another categorization is introduced in addition to the previous one: local vs global anomaly explanation methods. A local method explains one outlier at a time and a global method provides explanations for all the anomalies of the dataset at once. In [7], feature-based explanations, semantic explanations, visualisation techniques, metrics, model-specific methods and model-agnostic methods are used as categories of anomaly explanation methods. In [15], several categories are identified: anomaly detectors with explanations, outlier explanations for groups of outliers, outlying aspects mining (which identifies which subset of features makes a data point different from the rest of the dataset), outlying property detection (which finds the feature that makes a data point different from the data points that are the most similar to it), pictorial explanations, decision rules and sequential feature explanations. Although the work in [18] is mainly a survey on anomaly detection algorithms, the problem of anomaly explanation is also discussed. A distinction is made between model-agnostic explanation methods and neuralization which is the conversion of machine learning models into neural networks in order to use explanation methods developed for neural networks, whatever they are used for. In the other works related to the topic, anomaly explanation methods are listed without a particular classification. The coarsest taxonomies (local vs global and model-agnostic vs model-specific) do not take into account the specificities of the anomaly detection topic. The other taxonomies provide categories which are not really well-distinguishable: in [7] for instance, the feature-based explanations can be model-specific or model-agnostic; semantic explanations and visualisation techniques can be used for feature-based explanations.

3 Taxonomy of Anomaly Explanation Methods

We will consider the following example: in Table 1, we have a list of products along with their model, unit weight (W) and unit price (P). We want to identify

the anomalous products, using the information in the last two columns of the table which represent the true weight (TW) of each product and its true price range (TP) observed on online merchants.

Table 1: List of products and their true characteristics

ID	Model	W (g)	P (USD)	TW (g)	TP (USD)
1	iPhone X	174	550	174	[500-600]
2	iPhone 11	194	600	194	[800-1000]
3	iPhone 12	300	500	164	[1100-1500]
4	Galaxy S20	163	850	163	[800-900]
5	Galaxy S21	169	900	169	[900-1200]
6	Galaxy Note 20	250	900	192	[550-700]
7	MI 11	100	500	196	[450-600]
8	MI 10S	208	300	208	[100-350]
9	POCO F2 Pro	260	800	210	[200-300]

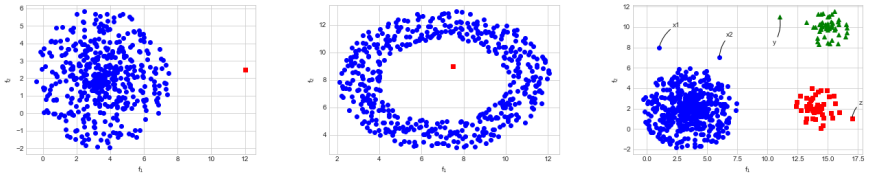
From the table, it can be seen that the anomalies are: the product *2* because of its low price, the product *3* because of its high weight and its low price, the products *6* and *9* because of their high weight and their high price, and the product *7* because of its low weight.

According to the reasons why a product is an anomaly, it is shown hereafter that four types of explanations may be envisaged: **explanation by feature importance**, **explanation by feature values**, **explanation by data points comparisons** and **data structure aware explanation**.

3.1 Anomaly Explanation By Feature Importance

For an algorithm which aims at recognizing in a set of images which ones are cat images and which ones are dog images, the most natural way to tell users why the algorithm tagged a picture as a cat instead of a dog is to return the group of pixels that helped the algorithm to make the difference. This group of pixels can represent the whiskers of the cat on each image for example. In this way, the user will notice that the whiskers are an attribute that the cat possesses, and not the dog, and will therefore understand why the algorithm decided that it is a cat picture. In general, identifying the features/attributes which contributed the most to the decision of an algorithm is a good start and a classical method to provide explanations. Anomaly detection is also concerned. In Figure 1a below, to mark the square data point as anomalous, we can look only at the feature f_1 for all the instances: in comparison to the regular data points in blue for which the values of the attribute f_1 vary between -1 and 8, it takes the value 12. The same cannot be told for the feature f_2 since the square instance has a value of 2.5 for that attribute, which is normal when compared to the values of f_2 for the regular instances. Consequently, to explain that anomaly to the user, it can be said that attribute f_1 contributed to the abnormality of

the square data point. This first category of anomaly explanation is **feature importance**. The contribution of each feature can be weighted or not. With weighted feature importance, the feature *unitprice* will receive a higher weight than the feature *unitweight* for outlier *9*. Both attributes contribute to making the instance anomalous, but the feature *unitprice* contributes the most because it is further away from the regular values than *unitweight* is, for that instance. This type of explanations is the most explored one. The works in [1], [4], [9] and [16] for example provide weighted feature importance explanations. Those in [5], [14] and [20] are non-weighted feature importance anomaly explanation methods. They just return the most important features without quantifying their priority with respect to anomaly identification.



(a) (b) (c)

Fig. 1: The necessity of anomaly explanation by feature importance (a) by feature values (b) and by structure analysis (c)

Anomaly explanation methods based on feature importance do not only provide information about why a specific data point is anomalous, but they can also give a global understanding of the anomalies by identifying the features that explain a set of anomalies or all the anomalies. Furthermore, feature importance can help identify different groups of anomalies, like in [16] where the authors propose a clustering of the anomalies based on the features gradients to identify the types of anomalies present in the data set. But if the original features are transformed prior to the anomaly detection, feature importance scores will not be meaningful to the final users as they will not recognize the features presented by the explanation system. In addition to that, just telling which features are important is sometimes not enough. In Figure 1b, when trying to explain the abnormality of the square data point using feature importance, we will observe that both features have equal importance, because one attribute does not help the algorithm to identify the anomaly more than the other: the isolated instance has a regular value for each of the features taken independently. It is the combination of the values for both attributes which makes the data point irregular. In this case, explanation by feature importance will return the attribute pair $\langle f_1, f_2 \rangle$, and that is little information since the complete attribute space is returned. In two dimensions it is easy for the user to plot and observe. But, if we are in higher dimension, which is almost always the case, displaying a list of features with more than two having the same importance is not really helping

the user. In these situations, it would have been more helpful to say, for instance, that the data point in Figure 1b is anomalous because it has a value for the feature f_1 around 7.5 and a value for the feature f_2 around 9. This second category of explanation is called **anomaly explanation by feature values**.

3.2 Anomaly Explanation By Feature Values

All the explanations coming from decision-tree-based anomaly detection algorithms lie in this category. Explanations are in the Disjunctive Normal Form (DNF), and each literal of the DNF is a conjunction of predicates. Each predicate is a condition on the value of a feature which has the form $f\ s\ v$ where f is a feature, s is one of the signs $<, \leq, =, >, \geq$ and v is a feature value. As an illustration, an explanation by feature values of outlier *9* can be: $unitweight \geq 210$ and $unitprice \geq 300$. Works like those in [2], [3], [10] and [22] belong to this category of anomaly explanation methods. Counterfactual explanations can also be classified in this category. Counterfactual explanations indicate which feature values to change (and how) in order to obtain a different prediction for an instance. For example, a counterfactual explanation of the outlier *2* will indicate that the unit price must be increased by 200 to obtain a regular instance. Counterfactual explanations in the context of anomaly detection are explored in [6].

The rules can easily become unreadable due to their number. As a result, some authors choose to return a short list of rules, each rule having a limited number of predicates. This can be sub-optimal because some less important (but still important) information about why an instance is anomalous may be ignored. Another flaw of this type of explanations is that, unlike feature importance, it is a bit complicated to explain anomalies globally. In addition to that, extracting and consolidating rules is more complex in terms of time processing. However, rules remain the most natural way of explaining anomalies, and translating rules to a pseudo natural language is relatively easy.

With the two previous categories of explanations, we just have information about the anomaly. We do not know concretely what is the difference between anomalies and regular data points. With the example in Figure 1b, after discovering that the instance is anomalous because $f_1 = 7.5$ and $f_2 = 9$, the user can ask if a data point with $f_1 = 8$ and $f_2 = 7$ would be anomalous (without plotting the data set of course). Explanations by feature importance and by feature values do not provide an answer to this question. An answer would be provided if the anomaly was explained by directly comparing it to regular data points. This third category of explanations will be called **anomaly explanation by data points comparison**.

3.3 Anomaly Explanation By Data Points Comparisons

Angle-Based Outlier Detection (ABOD) [11] is an unsupervised anomaly detection method providing explanations. To give explanations on why an instance is outlying, ABOD finds its closest instance in the nearest cluster, then computes

and returns the difference vector between the two data points. Works in [13], [17] and [21] also belong to this category.

Displaying similar instances and showing the differences between the anomalous instance and similar instances allow the user to concretely and easily perceive why a data point is irregular. But these explanations are very limited by the choice of a distance/similarity metric and require distances computation to find similar instances. Plus, this kind of explanation is not very informative when used alone. When used in combination with the two first categories of explanations, it can provide richer explanations to anomalies.

But if there are different clusters of regular data points in the data set, and each of these clusters has some anomalies as shown in Figure 1c where there are 3 clusters and 4 anomalies (x_1, x_2, y and z), the most complete explanation that can be provided is telling that x_1 and x_2 are anomalies for the cluster of round instances and why it is the case, that y is an anomaly for the the triangles and why, and finally that z is an anomaly for the squares and why.

To provide this kind of detailed explanations, an analysis of the intrinsic structure of the data set is required, followed by a comparison of the anomaly(ies) with this intrinsic structure. This last category of explanations will be called **explanation by structure analysis**. It starts at the anomaly detection level by identifying groups of anomalies and individual anomalies with respect to different groups of regular data points.

3.4 Anomaly Explanation By Structure Analysis

Analyzing the structure means discovering in the dataset groups of regular data points, groups of irregular data points, instances which deviate from each group and instances that are in groups where they are not supposed to be. In the example from table 1, products can be grouped according to the model in order to identify and explain the anomalies of each model. For example, outlier *2* is an outlier for the model *iPhone 12* because its price is lower than usual, for products of this model. An explanation by structure analysis should provide this information. Besides that, regular products can be grouped according to the true price range, in order to obtain different ranges of products. For example in 1, high-end products can be those which true prices range in the interval $[800 - 1500]$, low-end products those which true prices range from 100 to 400 and, an intermediate range of products can contain those for which $unitprice \in [450 - 700]$. With this breakdown, an explanation by structure analysis for the outlier *2* is that according to its unit price it is a mid-range product, but it is not normal because products of this model are supposed to be high-end products. This kind of explanations can be provided by analyzing in details (possibly manually) the detected anomalies, but the goal is to simplify the process as much as possible, for humans and for the computer. Identifying the anomalies and giving directly this type of detailed explanations could be very useful. Two works have been identified along these lines [12,19], but this type of explanation is sorely lacking references.

Anomaly explanation by structure analysis provides the most detailed information about why instances are anomalous and it is certainly the kind of explanation the most expected in various applicative contexts. But it has not been deeply explored yet. The works identified as belonging to this category are a sequence of steps (anomaly detection -> clustering -> analysis of the clusters). No method in the literature has provides so far a unified algorithm going directly from the detection to the detailed explanations. Also, the two methods identified in the literature explain anomalies in groups. But structure analysis should also be able to explain why a specific data point is anomalous, and not only why a set of instances are anomalous.

4 Conclusion

This work aimed at providing a categorization of anomaly explanation methods and at opening directions for future works on that crucial and topical field. Four categories were defined in order to provide a taxonomy which takes into account the particularities of anomaly detection and which is more refined than the taxonomies existing in the literature: **feature importance**, **feature values**, **data points comparisons** and **structure analysis**. Anomaly explanation by feature importance has been widely explored, in contrast to structure analysis which provides the most detailed explanations. For this last category of explanations, the integration of human experts can be investigated: a human expert can help describe the structure of the regularities or irregularities so as to facilitate the identification and the explanation of anomalies. In conclusion, a lot can still be done in relation to anomaly explanation. Although the field can leverage the methods developed for other tasks (like classification) or for neural networks, there is a need for explanation methods specifically built for anomalies.

References

1. Antwarg, L., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shap. arXiv preprint arXiv:1903.02407 (2019)
2. Barbado, A., Corcho, Ó., Benjamins, R.: Rule extraction in unsupervised anomaly detection for model explainability: Application to oneclass svm. arXiv preprint arXiv:1911.09315 (2019)
3. Baseman, E., Blanchard, S., DeBardeleben, N., Bonnie, A., Morrow, A.: Interpretable anomaly detection for monitoring of high performance computing systems. In: Outlier Definition, Detection, and Description on Demand Workshop at ACM SIGKDD. San Francisco (Aug 2016) (2016)
4. Carletti, M., Terzi, M., Susto, G.A.: Interpretable anomaly detection with diffi: Depth-based feature importance for the isolation forest. arXiv preprint arXiv:2007.11117 (2020)
5. Gupta, N., Eswaran, D., Shah, N., Akoglu, L., Faloutsos, C.: Beyond outlier detection: Lookout for pictorial explanation. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 122–138. Springer (2018)

6. Haldar, S., John, P.G., Saha, D.: Reliable counterfactual explanations for autoencoder based anomalies. In: 8th ACM IKDD CODS and 26th COMAD, pp. 83–91 (2021)
7. Hamelers, L.: Detecting and explaining potential financial fraud cases in invoice data with Machine Learning. Master's thesis, University of Twente (2021)
8. Hawkins, D.M.: Identification of outliers, vol. 11. Springer (1980)
9. Kauffmann, J., Müller, K.R., Montavon, G.: Towards explaining anomalies: a deep taylor decomposition of one-class models. Pattern Recognition **101**, 107198 (2020)
10. Kopp, M., Pevnỳ, T., Holeňa, M.: Anomaly explanation with random forests. Expert Systems with Applications **149**, 113187 (2020)
11. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 444–452 (2008)
12. Macha, M., Akoglu, L.: Explaining anomalies in groups with characterizing subspace rules. Data Mining and Knowledge Discovery **32**(5), 1444–1480 (2018)
13. Mejia-Lavalle, M.: Outlier detection with innovative explanation facility over a very large financial database. In: 2010 IEEE Electronics, Robotics and Automotive Mechanics Conference. pp. 23–27. IEEE (2010)
14. Micenková, B., Ng, R.T., Dang, X.H., Assent, I.: Explaining outliers by subspace separability. In: 2013 IEEE 13th international conference on data mining. pp. 518–527. IEEE (2013)
15. Mokoena, T.: Why is this an anomaly? Explaining anomalies using sequential explanations. Ph.D. thesis (2019)
16. Nguyen, Q.P., Lim, K.W., Divakaran, D.M., Low, K.H., Chan, M.C.: Gee: A gradient-based explainable variational autoencoder for network anomaly detection. In: 2019 IEEE Conference on Communications and Network Security (CNS). pp. 91–99. IEEE (2019)
17. Rieck, K., Laskov, P.: Visualization and explanation of payload-based anomaly detection. In: 2009 European Conference on Computer Network Defense. pp. 29–36. IEEE (2009)
18. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE (2021)
19. Shukla, A.K., Smits, G., Pivert, O., Lesot, M.J.: Explaining data regularities and anomalies. In: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–8. IEEE (2020)
20. Siddiqui, M.A., Fern, A., Dietterich, T.G., Wong, W.K.: Sequential feature explanations for anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) **13**(1), 1–22 (2019)
21. Smith-Renner, A., Rua, R., Colony, M.: Towards an explainable threat detection tool. In: IUI Workshops (2019)
22. Song, F., Diao, Y., Read, J., Stiegler, A., Bifet, A.: Exad: A system for explainable anomaly detection on big data traces. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 1435–1440. IEEE (2018)
23. Zhang, X., Marwah, M., Lee, I.t., Arlitt, M., Goldwasser, D.: Ace–an anomaly contribution explainer for cyber-security applications. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 1991–2000. IEEE (2019)