Beyond Numerical – MIXATON for outlier explanation on mixed-type data

Jakob Nonnenmacher
University of Oldenburg
Oldenburg, Germany
jakob.nonnenmacher@uol.de

Jorge Marx Gómez
University of Oldenburg
Oldenburg, Germany
jorge.marx.gomez@uol.de

Abstract— Outlier explanation approaches are employed to support analysts in investigating outliers, especially those detected by methods which are not intuitively interpretable such as deep learning or ensemble approaches. There have been several studies on outlier explanation in the last years. Nonetheless, there have been no outlier explanation approaches for mixed-type data. In this paper we propose multiple approaches for outlier explanation on mixed-type data. We benchmark them by using synthetic outlier datasets and by generating ground-truth explanation for real-world outlier datasets. The results on the various datasets show that while there is no approach that dominates others for all types of outliers and datasets, some can offer a consistently high performance.

Keywords-outlying aspect mining; outlier explanation; mixed-type; anomaly detection

I. INTRODUCTION

Being able to understand the output of a machine learning model is a core requirement for its successful deployment in real-world applications. Outlier detection models do not always fulfill this requirement. To remedy this, a number of outlier explanation methods have been proposed in the last years [1–5]. Outlier explanation is important because outlier detection methods are often used in a more explorative context which makes it important for analysts to investigate the results before they can take action. Furthermore, outlier detection methods often only provide information on whether something is an outlier but not why it is an outlier [1, 2, 5]. More recent state-of-the-art approaches based on deep-learning or ensemble methods are especially known for not being easily explainable [6]. This includes methods like IForest [7] as well as autoencoders [8]. Another aspect, which makes outlier explanation relevant, is the fact that all outlier detection methods employ some kind of statistical measure to determine the outliers. The problem with this is that the statistical measure might select an entry as an outlier that does not necessarily match the outlier definition within the domain the detector is used [4]. To weed out these potential false positives from the results, the analyst has to understand why an entry has been selected as an outlier.

Despite the large number of outlier explanation methods that have been proposed so far, most of them have been created for numerical data. A few methods explicitly address categorical data [9, 10] and one method has been presented which claims to work for both numerical as well as mixed-type data but has only been evaluated on the former [2]. Data in real-world applications is often mixed-type and multiple outlier detection methods and adaptions for mixed-type data exist [11–13]. Finding outliers, including those that are only apparent when considering both numerical and categorical features together, is important for uncovering fraud or identifying cyber-attacks. The fact that no outlier explanations approaches exist for mixed-type data makes the application of state-of-the-art mixed-type outlier detection approaches potentially less effective and practical for these and other real-world applications. To address this challenge, we propose and evaluate multiple variations of outlier explanation approaches for outliers detected on mixed-type data. Overall, the contribution of this work includes:

- Multiple adaptions of outlier explanation approaches for mixed-type data
- Creating a benchmark for outlier explanation approaches for mixed-type data
- Recommendations for which methods to use for mixed-type data

The rest of this paper is organized as follows: Sec. II introduces related work. Sec. III presents our developed approaches. Sec. IV contains the evaluation as well as the discussion of our results and in Sec. V we conclude our work.

II. RELATED WORK

The first prevalent outlier explanation approach is the so-called score-and-search approach in which subsets of the original full featurespace are created and then scored using a scoring metric. The subset featurespace in which an outlier receives the highest score is selected as the explaining subspace for that outlier. Multiple variations of this score-and-search approach have been proposed, focusing on reducing the number of subspaces to score [3, 4, 14–16] and on making the scores comparable between subspaces of different dimensionality [3, 15, 16].

The second prevalent approach is the so-called feature transformation and importance approach. Most of these approaches work by using a classifier that is trained to differentiate between the inliers and the outliers to explain. Then a feature selection technique like Lasso is used to determine the importance of the individual features for this classification task [5, 17–19]. One technique that is similar to this approach is to train a classification model and to then extract an explanation using the general model explanation technique SHAP which was

DOI reference number: 10.18293/SEKE2022-110

presented by [20]. SHAP is a method derived from game theory and attributes credit to each feature for the achieved prediction. In outlier explanation, this credit attribution is then used as the feature importance. This approach is, in combination with a classifier, used as a benchmark for the approaches presented by [2] and [1].

The approach by [1], called ATON, determines the feature importance by using a neural network with self-attention that learns to increase the distance between outlier and inlier. They use a triplet-loss by using the outlier to explain and sampling inliers from the neighborhood of that outlier as well as random inliers. After training the network, the self-attention is extracted from the network as the feature importance. They benchmark their approach against other numerical explanation approaches with their approach achieving superior performance.

The approach by [2], called Explainer, is an approach which works similar to isolation forests. It uses isolation trees which are built by performing splits in those leaves that contain the outlier. Leaves are split in such a way that the size of the resulting leaf containing the outlier is as small as possible [2]. At the end of the training, rules are extracted from the individual trees and the rules which are the most frequent among all trees are presented to the user to explain the outliers. The rule frequency is also used to determine the feature importance. This approach claims to work for mixed-type data but is not evaluated for it.

One prevalent way of differentiating outliers from inliers is their distance [21]. This is used in the approach by [1] when determining the local neighborhood of an outlier to select an inlier to differentiate it from. Since their method is designed for numerical data, they utilize simple Euclidean distance. However, the Euclidean distance might not be suitable when determining distances in mixed-type data [21].

Multiple approaches have been suggested to make mixed-type data usable in single-type methods. One-hot encoding is one of the most prevalent methods of allowing mixed-type data to be used in numerical methods [21]. In one-hot encoding, a set of binary dimensions is created for each categorical feature where each dimension stands for one unique feature value within the categorical features.

There are numerous distance metrics which have been designed for numerical data such as Euclidean or Manhattan distance [21]. Only a few have been designed for mixed-type data. One metric specifically designed for mixed-type data is the Gower distance [22].

III. Implementation

To develop an outlier explanation approach for mixed-type data, we used the currently best-performing method on numerical data, called ATON [1], as a foundation. We call the overarching family of our approaches *MIXATON*. We then designed multiple variations of this approach while incorporating specific adjustments for mixed-type data. The proposed methods are *MIXATON_OE_SUM*, *MIXATON_OE_AVG*, *MIXATON_GD* and *MIXATON_EL*.

1. *MIXATON_OE_SUM* This approach works by first encoding the categorical features using one-hot encoding and joining them with the numerical features. The neighborhood search for identifying inlier samples from the neighborhood of the outlier is then performed on this preprocessed dataset using k nearest neighbor with Euclidean distance. These, together with random inlier samples and the outlier are subsequently used to train the network. Afterwards, the learned self-attention is extracted from the network. To obtain the feature importance for the categorical features from the learned self-attention, the attention is added up for each created one-hot encoded feature per categorical feature. For the numerical features, the obtained attention can be used directly as importance. This way, one unique feature importance value is obtained per feature.

2. *MIXATON_OE_AVG* This variant is mostly equivalent to the *MIXATON_OE_SUM* approach. The key difference is that the attention for one-hot encoded categorical features is not summed but instead averaged to obtain the final importance of each categorical feature. This is done to prevent a potential overweighing of categorical features that could occur in the summing approach.

3. *MIXATON_GD* This approach addresses the potential unsuitability of using the Euclidean distance in the neighborhood search on one-hot encoded data for determining samples. This is done by using the Gower distance, as it is a distance metric specifically designed for mixed-type data, on the unprocessed dataset for determining the training samples. Only after determining the samples, the data is one-hot encoded and used for training the network.

4. *MIXATON_EL* The final variant we propose is the *MIXATON_EL* approach. In this approach, the categorical features are one-hot encoded in the beginning before the neighborhood search. To mitigate the effect of the resulting one-hot encoded features being seen as independent [23], embedding layers are used in this approach. The neural network is amended with one embedding layer for each categorical feature. The layer size is chosen for each categorical feature as half its cardinality. After the training of the network, the importance for the categorical features is averaged based on the resulting performance for each embedding layer.

To benchmark our developed variations, we used the only method which has been explicitly proposed for mixed-type data so far, the Explainer approach by [2] as well as XGBoost in combination with SHAP.

IV. Evaluation

In our evaluation, we first created suitable datasets on which we subsequently applied the different approaches.

A. Creating suitable datasets for evaluation

To be able to evaluate the explanation methods, we required both knowledge about which entries within the dataset are outlying as well as which features make these entries outlying. To achieve this, we followed two different approaches.

1) Synthetic outliers

For our first method, we adapt an approach which has been used for creating evaluation datasets in mixed-type outlier detection studies [11, 12]. In this approach datasets without obvious outliers are used and then artificial outliers are injected by shifting values in numerical features and swapping categories in categorical features. This way both the outliers as well as the responsible features are available as a ground-truth. The approach works by randomly selecting 10% of entries in the dataset. We then randomly selected 30% of the features of each of those entries. If the feature is numeric its value gets shifted by two times the feature's standard deviation. In cases in which the feature is categorical or binary the feature value is replaced by another value of that feature. The datasets we used are the Australian credit (a_credit), German credit (g_credit), Heart, Thoracic surgery (thoracic), Auto MPG and Contraceptive (contra) datasets from the UCI ML repository (https://archive.ics.uci.edu/ml/index.php).

2) Pseudo-ground-truth for real outliers

The advantage of using injected outliers is that it is objectively clear what caused the outliers but they might not accurately reflect what real outliers look like. To address this potential limitation, we created a second benchmarking dataset using datasets with real outliers, also from the UCI ML repository.

To generate pseudo-ground-truth explanations, all feature subspaces of these datasets were created and the outliers were scored using IForest, with one-hot encoded categorical features, as well as SPAD [24] and MIXMAD [12] as mixed-type outlier detection methods. Since it is not certain whether the used methods are dimensionally unbiased [16], the rank of the outlier in each subspace is used to determine the explaining subspace. This way, three ground-truth subspaces are obtained for each of the datasets. Since all subspaces had to be created for this approach, we only selected datasets with a dimensionality of 15 or less. This is done because with 16 dimensions or more over 65,000 subspaces would have to be scored using each method which would have been computationally infeasible for multiple datasets.

B. Conducting the evaluation

We evaluated our proposed approaches as well as the already existing method by [2], called *Explainer*, as well as *XGBoost* in combination with *SHAP* on the created datasets. We used multiple metrics to compare the methods' performance. All evaluated explanation methods return an ordered list of explaining features which we used, together with the ground-truth labels, to determine the performance. For providing a fair comparison with the *Explainer* method which only returns a limited subset of features, we employed R-Precision as a measure [25]. Using this metric, the Explainer method did not provide good results. The rigid way of splitting for specific feature values on categorical data might not be able to account for more complex outliers and thus leads to worse performance.

Since the *Explainer* method showcased the lowest performance and it does not return a ranking for the full featurespace, we are comparing the other methods on the complete featurespace using average precision (AP).The result for the synthetic outliers can be seen in TABLE I. When taking the whole feature set into account both the approach of summing the importance and the approach of averaging the importance seem to provide good results on the synthetic outliers.

TABLE I. AVERAGE PRECISION ON SYNTHETIC OUTLIERS

	XGBoost SHAP	MIXATON_ OE_SUM	MIXATON_ GD	MIXATON_ OE_AVG	MIXATON_ EL
a_credit	0.3928	0.4091	0.4153	**0.4193**	0.4047
autompg	0.4457	**0.4845**	0.4276	0.4468	0.4591
contra	0.5283	0.5191	0.4954	**0.5575**	0.4002
g_credit	0.4090	**0.4250**	0.4202	0.4237	0.4165
heart	0.3551	**0.4035**	0.3521	0.4022	0.4023
thoracic	0.3775	**0.4182**	0.4145	0.4098	0.3945
avg	0.4181	**0.4432**	0.4208	**0.4432**	0.4129

The result rated via the AP for the real outliers can be seen in TABLE II. All *MIXATON* approaches provide good performance with the method using the Gower distance achieved the best performance.

TABLE II. AVERAGE PRECISION ON REAL OUTLIERS

Dataset	XGBoost SHAP	MIXATON_ OE_SUM	MIXATON_ GD	MIXATON_ OE_AVG	MIXATON_ EL
abalone_iforest	0.4468	0.4218	0.4162	0.4309	**0.5271**
abalone_mixmad	0.5308	0.6089	0.5973	0.6308	**0.6382**
abalone_spad	0.3486	0.4694	0.4747	0.4577	**0.4996**
adap_iforest	**0.4091**	0.3486	0.3766	0.4109	0.3721
adap_mixmad	0.4067	**0.4320**	0.4285	0.3569	0.3961
adap_spad	**0.3465**	0.2734	0.3225	0.3400	0.2931
credit_iforest	0.4112	**0.4440**	0.4297	0.3894	0.3871
credit_mixmad	**0.4522**	0.4492	0.4333	0.4094	0.4115
credit_spad	0.3438	0.3863	**0.3858**	0.3526	0.3460
heart_iforest	0.3642	0.3982	**0.4096**	0.3881	0.3937
heart_mixmad	0.4445	0.4498	**0.4693**	0.4526	0.4616
heart_spad	0.3680	0.3923	**0.4069**	0.3529	0.3628
mammography_iforest	0.6389	0.6870	**0.7124**	0.6327	0.5444
mammography_mixmad	0.6995	0.7406	0.7254	**0.7426**	0.6492
mammography_spad	0.6483	0.6705	**0.6786**	0.6554	0.5889
avg	0.4573	0.4781	**0.4844**	0.4669	0.4581

Some of the key observations for the overall results are:

1. No single approach is always superior or inferior. This is to be expected since each data set has different characteristics.
2. Certain approaches provide consistently better performance on a large variety of ground truth labels and datasets. The *MIXATON_GD*, *MIXATON_OE_SUM* and

MIXATON_OE_AVG approaches provide the best performance. This is notable, since these approaches have been introduced for the first time in this paper.

3. There are some approaches that are stronger on certain datasets while other approaches are stronger on other datasets (XGBoost and SHAP, *MIXATON_GD*). This insight means that it might be possible to construct approaches that utilize the strength of multiple approaches to obtain a superior performance.
4. The result of methods might depend on how important either the numerical or categorical features are for making a certain entry an outlier. Thus, the method *MIXATON_OE_SUM* might perform better on a dataset in which the categorical features are more important for making an outlier outlying while *MIXATON_OE_AVG* might perform better when the responsibility is evenly distributed between the feature types. This should be further investigated in future research.

V. Conclusion

In this paper, we propose multiple approaches for explaining outliers on mixed-type data. We also perform the first benchmark of outlier explanation approaches on mixed-type data and describe two approaches for preparing suitable benchmark datasets for this purpose. We show that our proposed approaches *MIXATON_OE_SUM*, *MIXATON_OE_AVG* and *MIXATON_GD* show good performance on various datasets and outperform previously introduced approaches. Using these methods, outlier detection can be made more useful in domains in which it is applied on mixed-type data.

One important question that is raised in this comparison of multiple approaches is: *Which explanation method is the most suitable for explaining my detected outliers?* Our experimental findings suggest that there is no one best performing outlier explanation measure for all mixed-type datasets. Although, the different *MIXATON* variations we propose, apart from *MIXATON_EL*, all seem to provide consistently good performance on most datasets.

One possible approach to mitigate the fact that different methods perform better on different datasets would be to combine explanation methods in an ensemble approach. This should be investigated in future research.

References

[1] H. Xu *et al.*, "Beyond Outlier Detection: Outlier Interpretation by Attention-Guided Triplet Deviation Network," in *Proceedings of the Web Conference 2021*, 2021, pp. 1328–1339. [Online]. Available: https://doi.org/10.1145/3442381.3449868

[2] M. Kopp, T. Pevný, and M. Holeňa, "Anomaly explanation with random forests," *Expert Systems with Applications*, vol. 149, pp. 1–18, 2020, doi: 10.1016/j.eswa.2020.113187.

[3] D. Samariya, S. Aryal, K. M. Ting, and J. Ma, "A New Effective and Efficient Measure for Outlying Aspect Mining," in *Web Information Systems Engineering – WISE 2020*, 2020, pp. 463–474.

[4] M. A. Siddiqui, A. Fern, T. G. Dietterich, and W.-K. Wong, "Sequential Feature Explanations for Anomaly Detection," *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 1, Article 1, 2019, doi: 10.1145/3230666.

[5] N. Liu, D. Shin, and X. Hu, "Contextual outlier interpretation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden: AAAI Press, 2018, pp. 2461–2467.

[6] M. Carletti, M. Terzi, and G. A. Susto, *Interpretable Anomaly Detection with DIFFI: Depth-based Feature Importance for the Isolation Forest*.

[7] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation Forest," 2008, doi: 10.1109/ICDM.2008.17.

[8] J. Nonnenmacher, F. Kruse, G. Schumann, and J. Marx Gómez, "Using Autoencoders for Data-Driven Analysis in Internal Auditing," *HICSS*, 2021, doi: 10.24251/HICSS.2021.697.

[9] H. Xia, H. Q. Vu, J. Tan, X. Li, and G. Li, "Characterizing the Outlying Feature Set of Groups," *Procedia Computer Science*, vol. 165, pp. 119–125, 2019, doi: 10.1016/j.procs.2020.01.086.

[10] F. Angiulli, F. Fassetti, and L. Palopoli, "Detecting Outlying Properties of Exceptional Objects," *ACM Trans. Database Syst.*, vol. 34, no. 1, 2009, doi: 10.1145/1508857.1508864.

[11] S. Eduardo, A. Nazabal, C. K. I. Williams, and C. Sutton, "Robust Variational Autoencoders for Outlier Detection and Repair of Mixed-Type Data," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, S. Chiappa and R. Calandra, Eds., Proceedings of Machine Learning Research: PMLR, 2020, 4056--4066. [Online]. Available: http://proceedings.mlr.press/

[12] K. Do, T. Tran, and S. Venkatesh, "Energy-based anomaly detection for mixed data," *Knowledge and Information Systems*, vol. 57, no. 2, pp. 413–435, 2018, doi: 10.1007/s10115-018-1168-z.

[13] M. Garchery and M. Granitzer, "On the influence of categorical features in ranking anomalies using mixed data," *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia*, vol. 126, pp. 77–86, 2018, doi: 10.1016/j.procs.2018.07.211.

[14] J. Zhang and H. Wang, "Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance," *Knowledge and Information Systems*, vol. 10, no. 3, pp. 333–355, 2006, doi: 10.1007/s10115-006-0020-z.

[15] N. X. Vinh *et al.*, "Discovering outlying aspects in large datasets," *Data Mining and Knowledge Discovery*, vol. 30, no. 6, pp. 1520–1555, 2016, doi: 10.1007/s10618-016-0453-2.

[16] L. Duan, G. Tang, J. Pei, J. Bailey, A. Campbell, and C. Tang, "Mining outlying aspects on numeric data," *Data Mining and Knowledge Discovery*, vol. 29, no. 5, pp. 1116–1151, 2015, doi: 10.1007/s10618-014-0398-2.

[17] B. Micenková, R. T. Ng, X. Dang, and I. Assent, "Explaining Outliers by Subspace Separability," 2013, doi: 10.1109/ICDM.2013.132.

[18] X. H. Dang, B. Micenková, I. Assent, and R. T. Ng, "Local Outlier Detection with Interpretation," in *Machine Learning and Knowledge Discovery in Databases*, 2013, pp. 304–320.

[19] X. H. Dang, I. Assent, R. T. Ng, A. Zimek, and E. Schubert, "Discriminative features for identifying and interpreting outliers," *2014 IEEE 30th International Conference on Data Engineering*, 2014.

[20] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[21] C. C. Aggarwal, *Outlier analysis*: Springer, 2017.

[22] T. Wangchamhan, S. Chiewchanwattana, and K. Sunat, "Efficient algorithms based on the k-means and Chaotic League Championship Algorithm for numeric, categorical, and mixed-type data clustering," *Expert Systems with Applications*, vol. 90, pp. 146–167, 2017, doi: 10.1016/j.eswa.2017.08.004.

[23] P. Cerda and G. Varoquaux, "Encoding high-cardinality string categorical variables," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2020, doi: 10.1109/TKDE.2020.2992529.

[24] S. Aryal, K. M. Ting, and G. Haffari, "Revisiting Attribute Independence Assumption in Probabilistic Unsupervised Anomaly Detection," in *Intelligence and Security Informatics*, 2016, pp. 73–86.

[25] G. O. Campos *et al.*, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 891–927, 2016, doi: 10.1007/s10618-015-0444-8.