VIAL-AD: Visual Interactive Labelling for Anomaly Detection – An approach and open research questions

Andreas Theissler[1][0000−0003−0746−0424], Anna-Lena Kraft[2], Max Rudeck[1], Fabian Erlenbusch[1]

[1]Aalen University of Applied Sciences, 73430 Aalen, Germany
[2]IT-Designers Gruppe, 73730 Esslingen, Germany

Abstract. In anomaly detection problems the available data is often not or not fully labelled. This leads to results that are usually significantly worse than in balanced classification problems. In this short paper VIAL-AD is proposed, which addresses this problem with a sequence of unsupervised, semi-supervised and supervised machine learning models allowing a user to interactively label data points. This allows to move towards supervised anomaly detection, starting with unlabelled data. The approach is introduced and identified open research questions are discussed.

Keywords: visual interactive labelling · VIAL · anomaly detection · human-centered machine learning

1 Introduction

This work addresses machine learning-based anomaly detection (AD) [6, 1], where the aim is to classify data points as either *normal* or *anomaly* based on a set of features f. This can be achieved by training AD models on data that is (a) unlabelled, (b) contains labelled normal data, or (c) contains labelled normal data and anomalies. Applications of AD can be found in system health monitoring, intrusion detection, fraud detection, and the analysis of medical data. One main application field is data-driven fault detection, e.g. addressed in [20, 21].

This work is motivated by the question of *how we can compensate for the lack of a labelled and representative data set in AD problems by incorporating human knowledge in order to move to supervised AD.* While there are statistical or unsupervised ML methods to identify outliers, only a human expert can decide whether a data point is a true anomaly for a given application. Therefore, it suggests itself to incorporate the user in the process. We argue that for anomaly detection, this is indeed even more crucial than for balanced classification problems.

In this paper *visual interactive labelling for anomaly detection (VIAL-AD)* is proposed which – starting with unlabelled data – allows to iteratively move from unsupervised to supervised anomaly detection [6]. This is achieved by a

combination of (1) a sequence of machine learning (ML) models with different levels of supervision and (2) the incorporation of the user to interactively label data. The idea is to use unsupervised AD to address the so-called cold-start problem [22] in order to obtain an initial set of tentative labels. A sequence of AD models is used to suggest labels and a human expert confirms or overrules suggestions and labels data or regions in the feature space. We believe that in AD, where it is unlikely to have a representative and labelled training set, the user-in-the-loop is key to allow for the use of ML and move towards an accuracy that allows for productive use. In order to validate the idea, a prototype was implemented. Preliminary results are promising, however a number of open research questions were uncovered and are discussed in Section 3.

In the following, related work is briefly reviewed. Holzinger et al. showed how a user in-the-loop with ML models can improve the overall performance of a system [11]. A generic process for visual interactive labelling was proposed by Bernard et al. under the name of VIAL [4]. In [3] it was shown that VIAL can outperform pure active learning – specifically for two-class problems. In [2] AD models were used for interactive labelling, however not with the aim to label an AD data set. Trittenbach et al. discuss open research challenges for one-class active learning [22], e.g. the cold-start problem.

2 The approach: VIAL-AD

VIAL-AD consists of the steps *unsupervised*, *semi-supervised*, and *supervised AD* [6] (see Fig. 1(a)), where in each step model and user collaborate as follows: The model classifies the data and suggests the labels *normal* and *anomaly*. The user inspects the data points and their labels and (a) adjusts model hyperparameters, or (b) confirms/overrules the labels proposed by the model, or (c) visually labels data points or regions. Following that, the user decides to move to the follow-up step or to refine the labelling in the current step.

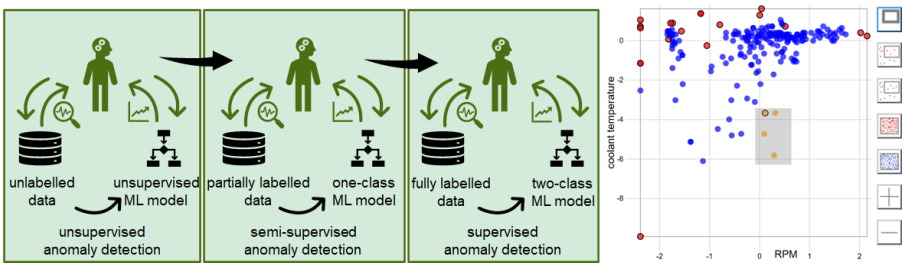

Fig. 1: (a) VIAL-AD: Interactive labelling with a sequence of ML models and the user-in-the-loop. (b) UI for visual interactive labelling, where the data shows RPM (revolutions per minute) and coolant temperature of vehicles during the occurrence of fault codes, read-out in car repair shops [19].

1. *unsupervised AD:* An unsupervised model (LOF [5] in current implementation) suggests initial labels which are confirmed or overruled by the user. Confirmed anomalies are moved to the two-class data set for the supervised step and are excluded for the current step. The model continues to report anomalies which are again evaluated by the user. In addition, regions in the feature space can be marked as *normal* or *anomaly* resulting in the creation of artificial data points. The result of the unsupervised step is a tentatively labelled train set.
2. *semi-supervised AD:* The reduced data set from the unsupervised step is used to train a one-class classifier (current prototype uses a OC-SVM [16]). For each data point the classifier suggests the labels *normal* or *anomaly*. These are processed analogously to the first step. A promising alternative is to make use of the anomalies from step 1 using methods like SVDDneg [18, 10]. This step's output is a labelled train set of normal data and anomalies, where the anomaly class is, however, not likely to be representative.
3. *supervised AD:* As VIAL-AD is used on real data, more and more anomalies are detected, so one can move to a supervised scenario. In addition to the normal data, the previously labelled anomalies are used to form a two-class train set that becomes increasingly representative. Hence, a variety of common ML models becomes applicable. In case of high class imbalance, sampling methods should be applied [12]. In the prototype ν-SVM is used.

The central element of interaction is an interactive 2D-scatter plot as in [2] (see Fig. 1(b)). This does however not limit the approach to two-dimensional data – higher dimensional data can be projected [15] onto two dimensions. Alternatively, visualisations for multi-dimensional data could be used. However, they introduce a higher complexity for the user.

3 Open research questions

Potential disruption caused by subsequent models: Different AD models have differing underlying assumptions [6, 18] about anomalies. Some work with probabilistic distributions, others with distances, densities (e.g. LOF [5]), reconstruction errors (autoencoders), or the adaption of the maximum-margin assumption to the one-class case (one-class SVMs [18, 16]). Hence, the use of different models in subsequent steps can induce disruptions in the way labels are suggested. A subsequent model could come up with a different labelling, which is confusing for the user. This disruption is to be minimised.

Visualisation-vs.-model dilemma: For data with > 2 dimensions the user is presented projected data, while the model may work on the original or on projected data. However, a projection with the aim to optimally visualise the data is not necessarily the optimal projection for the ML model to work on [14]. The original space or alternative projection methods might be more appropriate. Visualising and classifying different representations of the data can,

however, induce undesired effects. Wenskovitch et al. give an overview and potential solutions are discussed in [24].

Problem of non-interpretability of projected space: In AD problems, typically no representative set of anomalies exists. To compensate for that, entire regions could be marked as anomalous based on expert knowledge. In the original feature space these would be outlier values that can be clearly specified by experts. As discussed, the potentially high-dimensional data can be projected onto a lower dimensional space, the user can interact with. However, for many projection methods the relation between the visualization and the original input space is not obvious. This makes it difficult or even impossible for the user to label unoccupied regions in the feature space. Projection methods like t-SNE [23] aim to preserve the neighbourhood between data points, however do not preserve the properties in unpopulated regions. This creates a dilemma trying to mark unpopulated regions as *normal* or *anomaly*: in contrast to working on the original feature space, users do not have an intuition about where anomaly regions in the projected space are, as the projected feature space can be distorted and hardly interpretable [15]. This problem can be addressed in several ways:

1. Avoid projections by using visualization methods for high-dimensional data, which however makes interaction with the data more complex and does not scale well for a high number of dimensions.
2. Show original data objects for selected data points. Data types, where single data objects can be intuitively presented due to some order within the data are predestined for that, e.g. images, time series, or text. High-dimensional data in the form of independent feature vectors can, however, not easily be represented in an intuitive way.
3. Investigate projection methods and interaction facilities in order to allow for a user-friendly interaction [9, 8]
4. Let the user explore different projections, e.g. as proposed in [7].

Problem of highly imbalanced data: In AD, the distribution of the classes is typically highly imbalanced towards the *normal* class. As a consequence (a) this poses particular challenges for the projection methods, and (b) it raises the question if users will label the data accordingly.

In projections, anomalies should be positioned well separated from normal instances. Hence, an interesting issue is the sensitivity of projection methods to outliers. Bernard et al. evaluated different projection methods in [2]. While PCA's sensitivity to outliers is considered promising [2], in [3] it is stated that users prefer t-SNE [23]. The appropriateness of a projection method can be evaluated with a user study or using metrics specifying the readability of the projections. In [17, 13] ways to measure this readability are discussed.

The second question raised by the class imbalance is if users will label the data accordingly: On the one hand, users might run into the risk of overlooking anomalies due to their rareness. On the other hand, it might be that users overestimate the *anomaly* class, labelling too many data points as anomalous.

Risk of manual overfitting: Furthermore, an identified challenge is the risk of what we call "manual overfitting". In supervised ML, overfitting is addressed e.g. with regularization terms – preventing the model from too naively overfitting the data. However, with the user in-the-loop and with direct control over class labels, such a naive overfitting may take place: The user might be tempted to process the data set in such a way to achieve optimal accuracy as opposed to strictly applying domain knowledge to distinguish between normal or anomalous data points or regions in the feature space. This could be addressed with a – potentially high number – of *blind* test sets. Even after testing, this data should not be made available to the expert in order not to overfit towards the test set. Ideally, a test set should only be used once to evaluate performance. Another option would be the introduction of some regularization method, putting reasonable constraints on the user actions.

4 Conclusion

This paper discussed how the incorporation of humans can compensate for the lack of labelled data in anomaly detection. The proposed approach uses unsupervised AD on an initially unlabelled data set and lets the user confirm or overrule decisions. After having interactively processed the data set in collaboration with unsupervised AD models, the user can move to semi-supervised or supervised models. The key benefit of VIAL-AD is, that it allows to move towards supervised ML where it was previously not applicable due to the lack of labelled data. This is a problem often encountered in industry where data is recorded for a different purpose and the opportunities of applying ML are discovered later. While preliminary experiments indicate the applicability of VIAL-AD, open research questions were identified which will be addressed in future work. Following that, the goal is to evaluate VIAL-AD in a systematic user study and to apply it in a real-world case study.

References

1. An experimental evaluation of novelty detection methods. Neurocomputing **135**, 313 – 327 (2014)
2. Bernard, J., Dobermann, E., Sedlmair, M., Fellner, D.W.: Combining cluster and outlier analysis with visual analytics
3. Bernard, J., Hutter, M., Zeppelzauer, M., Fellner, D., Sedlmair, M.: Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study. IEEE Transactions on Visualization and Computer Graphics **24**(1), 298–308 (2018)
4. Bernard, J., Zeppelzauer, M., Sedlmair, M., Aigner, W.: VIAL: a unified process for visual interactive labeling. The Visual Computer **34**(9), 1189–1207 (2018)
5. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying Density-Based Local Outliers. In: SIGMOD Conference. pp. 93–104 (2000)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41**(3), 15:1–15:58 (Jul 2009)

7. Cutura, R., Holzer, S., Aupetit, M., Sedlmair, M.: VisCoDeR: A tool for visually comparing dimensionality reduction algorithms. In: 26th European Symposium on Artificial Neural Networks, ESANN 2018, Bruges, Belgium, April 25-27, 2018 (2018)
8. Endert, A., Fiaux, P., North, C.: Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2879–2888 (2012)
9. Faust, R., Glickenstein, D., Scheidegger, C.: DimReader: Axis lines that explain non-linear projections. IEEE Transactions on Visualization and Computer Graphics **25**(1), 481–490 (2019)
10. Görnitz, N., Kloft, M., Rieck, K., Brefeld, U.: Toward supervised anomaly detection. J. Artif. Int. Res. **46**(1), 235–262 (Jan 2013)
11. Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G.C., Pintea, C.M., Palade, V.: Interactive machine learning: experimental evidence for the human in the algorithmic loop. Applied Intelligence **49**(7), 2401–2414 (2019)
12. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence pp. 1–12 (2016)
13. Micallef, L., Palmas, G., Oulasvirta, A., Weinkauf, T.: Towards Perceptual Optimization of the Visual Design of Scatterplots. IEEE Transactions on Visualization and Computer Graphics **23**(6), 1588–1599 (2017)
14. Sacha, D., Sedlmair, M., Zhang, L., Lee, J.A., Peltonen, J., Weiskopf, D., North, S.C., Keim, D.A.: What you see is what you can change: Human-centered machine learning by interactive visualization. Neurocomputing **268**, 164–175 (2017)
15. Sacha, D., Zhang, L., Sedlmair, M., Lee, J.A., Peltonen, J., Weiskopf, D., North, S.C., Keim, D.A.: Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. IEEE Transactions on Visualization and Computer Graphics **23**(1), 241–250 (2017)
16. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation **13**, 1443–1471 (July 2001)
17. Sedlmair, M., Tatu, A., Munzner, T., Tory, M.: A taxonomy of visual cluster separation factors. Computer Graphics Forum **31**(3pt4), 1335–1344
18. Tax, D.M.: One-class classification. Concept-learning in the absence of counter-examples. Ph.D. thesis, Delft University of Technology (2001)
19. Theissler, A.: Multi-class novelty detection in diagnostic trouble codes from repair shops. In: 2017 IEEE 15th International Conference on Industrial Informatics (INDIN). pp. 1043–1049 (2017)
20. Theissler, A.: Detecting anomalies in multivariate time series from automotive systems. Ph.D. thesis, Brunel University London (2013)
21. Theissler, A.: Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. Knowledge-Based Systems **123**(C), 163–173 (May 2017)
22. Trittenbach, H., Englhardt, A., Böhm, K.: Validating one-class active learning with user studies – A prototype and open challenges
23. van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. Journal of Machine Learning Research **9**(11), 2579–2605 (2008)
24. Wenskovitch, J., Crandell, I., Ramakrishnan, N., House, L., Leman, S., North, C.: Towards a systematic combination of dimension reduction and clustering in visual analytics. IEEE Transactions on Visualization and Computer Graphics **24**(1), 131–141 (2018)