OCADaMi: One-Class Anomaly Detection and Data Mining toolbox *

Andreas Theissler[1], Stephan Frey, and Jens Ehlert

[1]Aalen University of Applied Sciences, Germany
andreas_theissler @ web.de
ORCID 0000-0003-0746-0424

Abstract. This paper introduces the modular anomaly detection toolbox OCADaMi that incorporates machine learning and visual analytics. The case often encountered in practice where no or only a non-representative number of anomalies exist beforehand is addressed, which is solved using one-class classification. Target users are developers, engineers, test engineers and operators of technical systems. The users can interactively analyse data and define workflows for the detection of anomalies and visualisation. There is a variety of application-domains, e.g. manufacturing or testing of automotive systems. The functioning of the system is shown for fault detection in real-world automotive data from road trials. A video is available: https://youtu.be/DylKkpLyfMk

Keywords: anomaly detection · machine learning · framework · visual analytics · demo

1 Introduction

This paper introduces the modular toolbox OCADaMi for one-class anomaly detection. Anomaly detection (AD) refers to reporting data points that have unexpected behaviour w.r.t. a training set, user experience or pre-defined thresholds. In this work machine learning-based approaches are used. Target users are developers, engineers, test engineers, and operators of technical systems. The application-domain is manifold, e.g. manufacturing or automotive systems.

In practice, a representative set of anomalies can often not be obtained. An example is fault detection, where a representative training set would mean to know all potential faults and to have corresponding labelled data. In contrast, normal data can easily be obtained from a system in normal operation mode. As opposed to using a two-class approach, an alternative is to use a training set of normal data and classify deviations as anomalies referred to as one-class classification [5]. OCADaMi can be used for:

1. interactive data analysis using visual analytics

* We thank IT-Designers GmbH and STZ Softwaretechnik for funding this research and many former associates and students for their contributions.

2. the creation of AD-workflows that can be run, optimized and exported
3. AD using the toolbox or by integrating the workflow in own applications
4. moving towards a supervised scenario, since during operation anomalies are detected gradually improving knowledge about the previously unknown or non-representative anomaly class

1.1 Related work

The application of AD on automotive data was shown in [3], in sensor networks in [1] and for intrusion detection systems in [4]. The authors have previously used OCADaMi for fault detection in automotive time series [9] and for the analysis of automotive after sales data [7].

In terms of alternative tools, Dd_tools by David Tax [6] has inspired the development of OCADaMi. While [6] is a feature-rich Matlab toolbox requiring the user to program, OCADaMi offers a configurable, user-centric framework with tightly integrated visualizations. In contrast to existing applications like KNIME or Rapid Miner, OCADaMi focusses on the specific problem of anomaly detection based on a training set of normal data and is not meant to be a generic data mining toolbox.

2 The anomaly detection toolbox OCADaMi

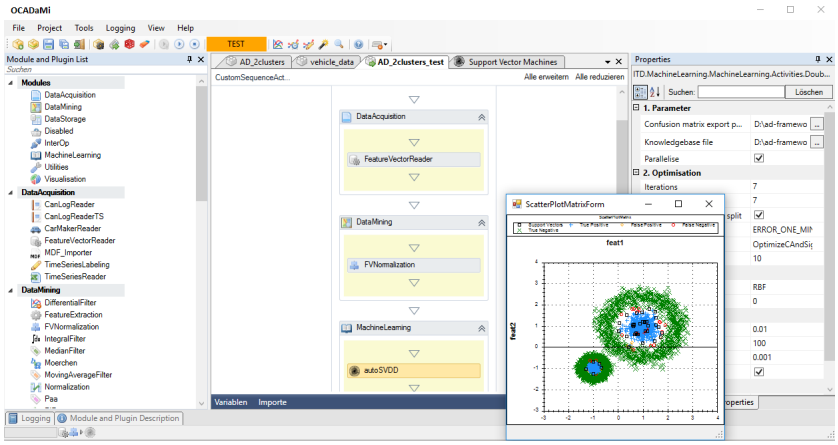

Fig. 1. Workflow with data acquisition, feature scaling and SVDD. Results are shown in the two-dimensional feature space.

OCADaMi is implemented as a modular, extensible framework in .NET, in specific in C#. The UI is implemented with Workflow foundation, Windows Presentation Foundation and WinForms and enables users to create sequential

workflows by chaining the desired plug-ins to fit their custom analysis needs (see 1). OCADaMi provides techniques for supporting the user in managing the workflows and in easily reproducing the results. Detailed reports are exported with results, in- and output of plug-in, generated visualisations, and knowledgebases. Advanced visualisation plug-ins are provided. As OCADaMi is implemented as a plug-in architecture, it is easily extensible by own plug-ins. The toolbox can interop with the R environments and exchange data. The toolbox is described following the steps of CRISP-DM [10]:

Data understanding: Using advanced interactive visualisations, value ranges, outliers, and correlations in feature space can be inspected.

Data preparation: As a form of user-centric machine learning, the application allows to incorporate expert knowledge to compensate the lack of a representative and labelled two-class training set. Visual analytics enables the user to filter the data, e.g. to remove measurement errors. Different pre-processing steps like scaling, time series filtering or resampling can be applied and time series data can be transformed to alternative representations.

Modelling: Users can select from a range of one-class models like a thresholded variant of k-NN as described in [8], LOF [2], and the one-class SVM support vector data description (SVDD) [5].

Evaluation: Results can be evaluated visually or on the basis of metrics.

Deployment: The toolbox can be used or the workflow can be integrated into own applications.

3 Case studies

Two brief case studies are shown here, both using one-class classification, i.e. exclusively training on normal data points. More details can be found in the accompanying video and in previous publications [9, 8].

3.1 Case study 1: Two banana shaped clusters

To show the functionality of OCADaMi, an artificial two-dimensional data set is used with two banana-shaped clusters with Gaussian noise. The test set is shown in 2a) where the right cluster corresponds to the normal class that was used to train the one-class classifier. This data set simulates the typical situation where anomalies are data points outside the normal value range in one or more features.

A thresholded variant of k-NN, as described in [8], is used for anomaly detection. Classification results on the blind test set are $TPR = 96.3\%$ and $TNR = 98.0\%$, where anomalies correspond to the negative class.

3.2 Case study 2: Real-world data from automotive tests

Anomaly detection in automotive systems is highly relevant due to the increasing complexity of modern vehicles, e.g. induced by safety or comfort systems. The

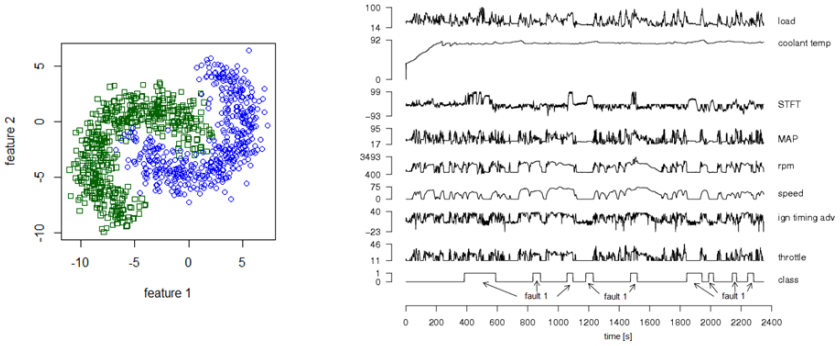

Fig. 2. Test sets of the case studies with normal data points and anomalies. a) Case study 1: Artificial data set with banana-shaped clusters where the right cluster is the normal class. b) Case study 2: Multivariate time series showing a recording from a vehicle road trial. Anomalies are depicted by the class label on the bottom.

training data consists of recordings from test drives in normal operation mode recorded during overland drives. The recordings are multivariate time series with 8 signals, like engine load, vehicle speed, and rpm. An example of one road trial in the test set is shown in 2b).

The focus is the detection of intermittent faults which manifest themselves in anomalies in the correlation of the signals. This is an imbalanced classification problem with a training set consisting of multiple normal multivariate time series with a total 24604 seconds at a sample rate of time point per second. The test set holds 12076 data points, with about 18% anomalies.

The data is first scaled, then SVDD with an RBF kernel is used. The toolbox offers autoSVDD, which autonomously determines the hyperparameters C and σ from the training set using recursive grid-search with repeated 7-fold cross validation. The determined hyperparameters are $C = 0.7778$ and $\sigma = 0.6968$. For more details the reader is referred to [8].

In a post-processing step subsequences are formed grouping together adjacent classified data points of the input time series. Since faults in the recordings can be of arbitrary length, variable-length subsequences are formed. Results on the blind test set from overland drives are $TPR = 78.7\%$ and $TNR = 73.8\%$. These results can be improved by building ensembles of multiple one- and two-class classifiers [9].

4 Conclusion

It was shown how one-class anomaly detection can be achieved using OCADaMi. Usage is easy and customizable due to its modular structure and the out-of-the-box implementations. Large data sets can be processed by transmitting the workflow queue to another computing node. Case studies showed AD on (1) artificial data sets and for (2) real-world data from automotive test drives.

References

1. Bosman, H.H., Iacca, G., Tejada, A., Wörtche, H.J., Liotta, A.: Ensembles of incremental learners to detect anomalies in ad hoc sensor networks. Ad Hoc Networks **35**(C), 14–36 (Dec 2015)
2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying Density-Based Local Outliers. In: SIGMOD Conference. pp. 93–104 (2000)
3. Prytz, R., Nowaczyk, S., Roegnvaldsson, T.S., Byttner, S.: Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. Engineering Applications of Artificial Intelligence **41**, 139–150 (2015)
4. Soudi, A., Khreich, W., Hamou-Lhadj, A.: An anomaly detection system based on ensemble of detectors with effective pruning techniques. In: IEEE International Conference on Software Quality, Reliability and Security. pp. 109–118. IEEE Computer Society (2015)
5. Tax, D., Duin, R.: Support vector data description. Machine Learning **54**(1), 45–66 (Jan 2004)
6. Tax, D.: Ddtools, the data description toolbox for matlab (Jan 2018), version 2.1.3
7. 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 (July 2017). https://doi.org/10.1109/INDIN.2017.8104917
8. Theissler, A.: Detecting anomalies in multivariate time series from automotive systems. Ph.D. thesis, Brunel University London (2013)
9. Theissler, A.: Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. Knowledge-Based Systems **123**(C), 163–173 (May 2017). https://doi.org/10.1016/j.knosys.2017.02.023
10. Wirth, R.: CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. pp. 29–39 (2000)