Ref. Ares(2019)6684260 - 29/10/2019

COORDINATED PRODUCTION
FOR BETTER RESOURCE EFFICIENCY

D2.3 Prototype software for anomaly detection and report on application results

divis - Germany

October 2019
www.spire2030.eu/copro

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 723575

Project Details

PROJECT TITLE	Improved energy and resource efficiency by better coordination of production in the process industries
PROJECT ACRONYM	**COPRO**
GRANT AGREEMENT NO	**723575**
INSTRUMENT	**RESEARCH AND INNOVATION ACTION**
CALL	**H2020-SPIRE-02-2016**
STARTING DATE OF PROJECT	**NOVEMBER, 1ST 2016**
PROJECT DURATION	**42 MONTHS**
PROJECT COORDINATOR (ORGANIZATION)	**PROF. SEBASTIAN ENGELL (TUDO)**

THE COPRO PROJECT

The goal of CoPro is to develop and to demonstrate methods and tools for process monitoring and optimal dynamic planning, scheduling and control of plants, industrial sites and clusters under dynamic market conditions. CoPro pays special attention to the role of operators and managers in plant-wide control solutions and to the deployment of advanced solutions in industrial sites with a heterogeneous IT environment. As the effort required for the development and maintenance of accurate plant models is the bottleneck for the development and long-term operation of advanced control and scheduling solutions, CoPro will develop methods for efficient modelling and for model quality monitoring and model adaption.

The CoPro Consortium

Participant No	Participant organisation name	Country	Organisation
1 (Coordinator)	Technische Universität Dortmund (TUDO)	DE	HES
2	INEOS Manufacturing Deutschland GmbH (INEOS)	DE	IND
3	Covestro Deutschland AG (COV)	DE	IND
4	Procter & Gamble Services Company NV (P&G)	BE	IND
5	Lenzing Aktiengesellschaft (LENZING)	AU	IND
6	Frinsa del Noroeste S.A. (Frinsa)	ES	IND
7	Universidad de Valladolid (UVA)	ES	HES
8	École Polytechnique Féderale de Lausanne (EPFL)	CH	HES
9	Ethniko Kentro Erevnas Kai Technologikis Anaptyxis (CERTH)	GR	RES
10	IIM-CSIC (CSIC)	ES	RES
11	LeiKon GmbH (LEIKON)	DE	SME
12	Process Systems Enterprise LTD (PSE)	UK	SME
13	Divis Intelligent Solutions GmbH (divis)	DE	SME
14	Argent & Waugh Ltd. (Sabisu)	UK	SME
15	ASM Soft S.L (ASM)	ES	SME
16	ORSOFT GmbH (ORS)	DE	SME
17	Inno TSD (inno)	FR	SME

Document details

DELIVERABLE TYPE	**PRELIMINARY REPORT**	
DELIVERABLE NO	**D2.3**	
DELIVERABLE TITLE	**PROTOTYPE SOFTWARE FOR ANOMALY DETECTION AND REPORT ON APPLICATION RESULTS**	
NAME OF LEAD PARTNER FOR THIS DELIVERABLE	divis	
LIST OF AUTHORS - NAME(S) AND ORGANISATION(S)	Peter Krause (divis)	
	Thomas Bäck (divis)	
	Christophe Foussette (divis)	
VERSION	V0.4	
CONTRACTUAL DELIVERY DATE	**30 OCTOBER 2019**	
ACTUAL DELIVERY DATE	**29 OCTOBER 2019**	
Dissemination level		
PU	Public	**X**
CO	Confidential, only for members of the consortium (including the Commission)	

Abstract

The goal of deliverable 2.3 is to add anomaly detection methods (see D2.2) to the software package ClearVu Analytics. Therefore, the command-line tool of ClearVu Analytics has been extended to allow a user to check univariate time series for potential anomalies. The anomaly detection method used is based on training regression models and using a statistical test with the distribution of the residuals of the predictions to find potential anomalies.

REVISION HISTORY

The following table describes the main changes done in the document since it was created.

Revision	Date	Description	Author (Organisation)
V0.1	20-09-2019	Creation and internal revision	Peter KRAUSE (divis)
V0.2	22-09-2019	Update	Thomas BÄCK (divis)
V0.3	23-09-2019	Review by TUDO	Simon WENZEL (TUDO)
V0.4	25-09-2019	Update in response to review	Peter KRAUSE (divis)
V1.0	03-10-2019	Approval by coordinator	Sebastian ENGELL (TUDO)

Disclaimer

THIS DOCUMENT IS PROVIDED "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION OR SAMPLE. Any liability, including liability for infringement of any proprietary rights, relating to use of information in this document is disclaimed. No license, express or implied, by estoppels or otherwise, to any intellectual property rights are granted herein. The members of the project CoPro do not accept any liability for actions or omissions of CoPro members or third parties and disclaims any obligation to enforce the use of this document. This document is subject to change without notice.

Table of contents

1 Executive summary

In the industrial environment, especially in the process industries, data collection and analysis become more and more important. One challenge in the data collection process is the occurrence of rare process states or corrupted data, which have a severe impact on the data analysis. These data points are commonly called outliers or anomalies.

This report sums up the progress on integrating anomaly detection methods into the software package ClearVu Analytics. In the previous work, anomaly detection methods have been compared and the most promising one (a regression based statistical comparison of residuals using an extreme studentised deviation test) has been integrated into ClearVu Analytics. The basis for the selection has been set in previous work as described in D2.2.

First, we describe the algorithm which has been implemented. Next, we show how it was integrated into ClearVu Analytics and how it can be applied to anomaly detection tasks. The effectiveness of the method has been validated using well known examples like the Yahoo data set. A test on a use case of this project has not yet been carried out because of insufficient data.

2 Anomaly Detection Method

One of the main tasks in anomaly detection is to find anomalies in a given univariate time series. In order to achieve such a task, the following method has been developed.

The main idea is to train a regression model on the time series which is able to predict the next time point depending on the time points in a small fixed time window w before the new time point. This implies that the model learns certain patterns in the time series. Thus, an anomaly can be detected if the predicted value is significantly different from the observed one. It is possible to detect collective anomalies if the small input pattern has not been observed multiple times and therefore the prediction is expected to deviate from the real observation. Because of the limitation of the window size, deviations in pattern length which exceed the input window cannot be detected.

After selecting a window size w a model is trained and is used to reconstruct the time series. The residuals $R = \{r_i\}_{i=w+1}^{n} = \{y_i - \hat{y}_i\}_{i=w+1}^{n}$ are then collected to find anomalies. Here, n is the number of data points, $\hat{y}_i$denotes the model prediction and y_i the observed value. The generalised extreme studentised deviate (ESD) test is applied to detect outliers in R. The hypotheses of ESD [1] are:

H_0: There are no anomalies in the data.

H_α: There are up to k percent of anomalies.

In ESD, only an upper bound is required for the suspected number of outliers and not the exact number of outliers. The test is performed in an iterative manner:

In each step, the extreme value $r^* = \arg\max_{r_i \in R} |r_i - \bar{r}|$ with sample mean $\bar{r}$ is checked against the null hypothesis. The test statistic is:

$$T = \frac{\max\{|r_i - \bar{r}| : r_i \in R\}}{s} = \frac{r^*}{s}$$

where s is the sample standard deviation. At the significance level α, the test statistics T is tested against the critical value:

$$\lambda = \frac{(n-1)t_{p,n-2}}{n\sqrt{n-2+t_{p,n-2}^2}}, p = 1 - \frac{\alpha}{2n}, n = |R|$$

with n denoting the size of the data set and $t_{p,n-2}$ denoting the $100 - p$ percentage point of the t-distribution with $n-2$ degrees of freedom [1]. The null hypothesis H_0 is rejected when $T > \lambda$. Regardless of the decision on H_0, the current extreme value r^* is removed from R and the same procedure is repeated on the remaining $n' = n - 1$ data points. The iteration stops when at most k percent of the data points have been removed from the initial R. Outliers are the removed data points on which the null hypothesis is rejected. Algorithm 1 shows the pseudo-code of the ESD algorithm.

Algorithm 1.Algorithm 1: Pseudo-code of the ESD algorithm

```
procedure ESD(k, R)
    n = |R|,  N = ⌈k · n⌉,  q = 0
    for i = 1, 2, ⋯, N do
        calculate r_i^* and λ_i
        D = D \ {r_i^*}
        n = n − 1
        if r_i^* / stdev(t) > λ then q = i
    next
    return {r_1^*, r_2^*, ⋯, r_q^*}
end
```

This approach has been tested against other approaches and overall good performance of this approach has been published in [2].

3 Implementation

ClearVu Analytics supplies a graphical user interface as well as a command-line tool for carrying out various tasks in the field of data analytics. As the main field of applications of anomaly detection is a recurrent task and needs to be integrated into existing workflows, we decided to implement the prototype of the anomaly detection into the command-line tool.

In order to integrate the anomaly detection into ClearVu Analytics the following steps have been carried out:

- Transform the univariate time series into a data matrix as a basis to train a regression model.
- Train different models and select one for the given task.
- Apply the ESD algorithm and present the results.

The transformation has been implemented and is done automatically prior to the modelling step. In order to train the models, the core functionality of ClearVu Analytics is connected to the anomaly detection task. Different modelling approaches can be trained simultaneously. For each approach the

hyperparameters of the modelling algorithm are optimised. The objective of the optimisation is the minimisation of the error on the validation parts of a cross-validation. The final models are compared and the most generalising model is selected to be used to generate the sets of residuals for the ESD algorithm. The ESD algorithm has been integrated into ClearVu Analytics producing the desired results which are available after running the command-line tool.

4 Usage

The command-line tool of ClearVu Analytics is available for Windows and Linux. It is called with a control file that contains all necessary information to carry out a given task [3]. A control file is a xml file. The control file for an anomaly detection task looks as follows:

```
<?xml version="1.0" encoding="utf-8" ?>
<Anomaly name="yahoo_r_1">
	<DataObject>
		<File name="real_1.csv" type="txt" header="true”/>
	</DataObject>
	<VariableSetObject>
		<VariableObject name="value" type="real" min="0.0" max="1.0" />
	</VariableSetObject>
	<ModelObject name="rf" output="Vplus1" type="RRandomForestModel">
		<TaskParameter name="DataSplit">NuTenTimesCross</TaskParameter>
		<TaskParameter name="DoOptimize">TRUE</TaskParameter>
		<ModelParameter name="ntree">500</ModelParameter>
	</ModelObject>
	<ModelObject name="mlp" output="Vplus1" type="MLPModel">
		<TaskParameter name="DataSplit">NuTenTimesCross</TaskParameter>
		<TaskParameter name="DoOptimize">TRUE</TaskParameter>
	</ModelObject>
	<GlobalParameters>
		<OutFile name="Yahoo_R1_Result.xml" exportCSV="TRUE"/>
		<TaskParameter name="window">25</TaskParameter>
		<TaskParameter name="k">10</TaskParameter>
		<TaskParameter name="alpha">0.05</TaskParameter>
	</GlobalParameters>
</Anomaly>
```

The first tag, which needs to be set, is **`Anomaly`** in order to set the task. Next, we need to set the path to the univariate time series which is done in the **`DataObject`** section. The user can choose which modelling algorithms to use. Each algorithm, which is intended to be used, is added by a **`ModelObject`** section. The details to set up such sections are described in the ClearVu Analytics manual. The anomaly detection algorithm itself has some parameters which can be set in the **`GlobalParameters`** section. The parameters are:

- **`window`**: The size of the window of values which are used to predict the next value. The default is 25.
- **`k`**: The maximal number of anomalies to look for. The default is 5.
- **`alpha`**: The confidence level. The default is 0.05.

The results are stored in a xml file (**`OutFile`**) and in a csv file. The xml file contains information about the models regarding their quality. This file can be imported into ClearVu Analytics for further analysis of the models. The csv file contains a series of zeros and ones labelling each time point as no anomaly (0) or potential anomaly (1). The header of the file contains the name of the model finally used.

This enables a user to integrate an anomaly detection task into an existing workflow by storing time series into a defined place and calling the command-line tool with a template of an anomaly task detection control file. The results can be further processed as needed depending on the overall task.

5 Summary and Outlook

This report documents the efforts to integrate anomaly detection into ClearVu Analytics. The first step was to decide which method to implement first. Based on the research done before, we decided to implement a method based on training regression models and applying an extreme studentised deviate test to find anomalies. We integrated this into ClearVu Analytics in a way that users can integrate this task into their existing workflows easily.

The implementation has been tested on academic examples but it was not possible to test it within CoPro use cases so far as the available data is not sufficient. We plan to test the implementation as soon as the data is available.

Depending on the acceptance of the users it is planned to add further methods for anomaly detection and to add a module to the graphical user interface to give users the ability to analyse results even more easily.

6 References

[1] B. Rosner, “Percentage Points for a Generalized ESD Many- Outlier Procedure,” *TECHNOMETRIC,* pp. 165-172, May 1983.

[2] S. Däubener, S. Schmitt, H. Wang, P. Krause and T. Bäck, “Anomaly Detection in Univariate Time Series: An Empirical Comparison of Machine Learning Algorithms,” in *19th Industrial Conference on Data Mining (ICDM 2019)*, New York, USA, 2019.

[3] divis intelligent solutions GmbH, *ClearVu Analytics User Manual,* Dortmund, 2019.