Automatic vs. Manual Feature Engineering for Anomaly Detection of Drinking-Water Quality

Valerie Fehst, Huu Chuong La, Tri-Duc Nghiem, Ben E. Mayer, Paul Englert, Karl-Heinz Fiebig
idatase GmbH
Frankfurt am Main, Hessen
info@idatase.de

ABSTRACT

This paper evaluates anomaly detection approaches for drinking-water quality. Two major machine learning techniques are compared. One is manual feature engineering with feature subset selection for dimensionality reduction. The other is automatic feature learning through a recurrent neural network. Both methods incorporate the time domain for change detection. Preliminary results show a superior performance of automatic feature learning with an F1 score of 80%. While the feature set proposed in this work outperforms naive classification with original features, it needs further analysis to reach comparable performance to the automatic approach.

CCS CONCEPTS

• **Information systems** → **Data analytics**; • **Computing methodologies** → Uncertainty quantification;

KEYWORDS

Anomaly detection, machine learning, fresh water, internet of things, deep learning

ACM Reference format:
Valerie Fehst, Huu Chuong La, Tri-Duc Nghiem, Ben E. Mayer, Paul Englert, Karl-Heinz Fiebig. 2018. Automatic vs. Manual Feature Engineering for Anomaly Detection of Drinking-Water Quality. In *Proceedings of Genetic and Evolutionary Computation Conference Companion, Kyoto, Japan, July 15–19, 2018 (GECCO '18 Companion),* 2 pages.
https://doi.org/10.1145/3205651.3208204

1 INTRODUCTION

Precise detection of changes in water quality is a crucial task for public water companies and urgently required for a fast reaction to contaminated drinking water. This problem motivates the investigation of machine learning methods for detecting changes and anomalies in water quality. The data set of the GECCO Challenge 2018 contains the time stamps, nine time series of various measurable water properties and a target vector of marked anomalies. Six of these measured features are quality indicators, including the pH value, Redox potential, electric conductivity, turbidity of water and the amount of chlorine dioxide in the two different water lines. Drastic changes of the behaviour in one of these features indicate anomalous events. While the amount of chlorine dioxide is controllable by humans, the other features are produced by external circumstances. The remaining three time series correspond to features that do not directly influence the water quality. These series are the water temperature and the flow rates of both water lines. Finally, a target vector marks time points in the series as abnormal or normal. This vector serves as labels for the anomaly detection algorithms in this work.

The results of the GECCO 2017 challenge shows highest performance on a second order polynomial feature space transformation. This indicates that the problem setting has a non-linear character and suggests the usage of non-linear anomaly detection methods. In order to create an appropriate bias-variance trade-off, the anomaly detection in drinking-water is systematically approached through data preprocessing and feature engineering. These methods allow to control the complexity of the model and obtain a decent set of parameters. Feature engineering is performed in two ways: First, a set of manual features incorporating the time domain is designed and statistical tests are employed to select a subset for detector training. Second, deep learning with a recurrent neural network is used to learn complex feature representations and detection models at the same time. The performance and decision behaviour of the detectors are assessed with the F1 scores. The following section describes the used method, the subsequent presents preliminary results and the last discusses these results and gives an outlook.

2 METHOD

2.1 Data Preprocessing

The first step when working with industrial data is the assurance of data quality. Missing values need to be either removed from the training set or interpolated. This work uses forward propagation for interpolation. It propagates the last valid observation forward up to the next valid one. This interpolation is suitable w.r.t. online operation in real-time settings, where only the last valid value is known at processing time. Furthermore, interpolated values do not fall outside the value range, thus preventing the creation of a false anomaly.

2.2 Manual Feature Engineering

This step aims to derive new, time-dependent features from the data as change detection occurs naturally in the time domain. Therefore, some time-dependent features are expected to better capture certain properties of the studied time series. For them, the first

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/10.1145/3205651.3208204

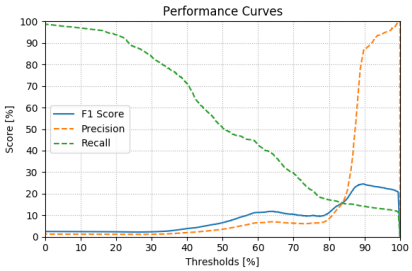

Figure 1: Logistic Regression on original features, the F1 score is 25%.

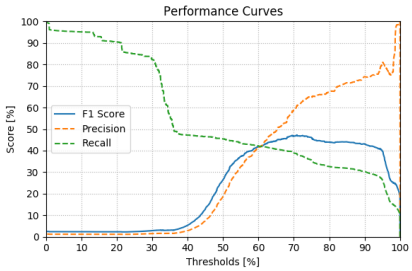

Figure 2: Logistic Regression on engineered features, the F1 score is 48%.

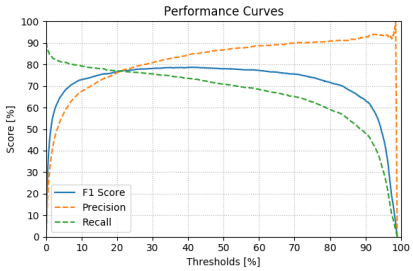

Figure 3: LSTM Neural Network with automatic feature learning, the F1 score is 80%

time-dependent feature is the lag operator c. It indicates how many preceding time steps are incorporated into the feature space.

The second time-dependent feature is derived by integrating the deviation between signal and reference points over a time interval. It takes the rate of change and the time duration into account. The time series is transformed via quadrature over an interval containing two time steps (i.e. $c = 2$).

The third time-dependent feature measures the complexity of the probability distribution [1] . Consider a sliding window, each composed of c time steps over which a histogram with b bins is constructed. Subsequently, the entropy over the probability of finding the time step x_t in the i-th bin is calculated through:
$H(x) = -\sum_{i=1}^{b} p_i \cdot \log p_i$.

The last time-dependent feature is the composition gradient. It measures the amount of change between the current state x_t and the previous state $f_{CG}(x_t; 1) = (x_t - x_{t-c})/c$ One of the binary classes is a minority and hence reduces the effective amount of data points required to fit models without overfitting. Dimensionality reduction of the feature space is therefore performed. The derived features are submitted to two feature selection criteria in order to determine a subset of features with high predictive power.

These criteria are the ANOVA F-value and the mutual information. The first determines the ratio between the variance of the features and the dependent variable. The latter measures the mutual information between features and dependent variable. The intersection of the k highest scoring features in both criteria constitutes the selected feature subset to span a k-dimensional feature space for an anomaly classifier.

2.3 Automatic Feature Learning

An alternative to the feature engineering is deep learning neural networks. Assuming that the data situation is large enough, neural networks render manual feature engineering obsolete by learn appropriate feature representations in their hidden layers. Detecting changes in drinking-water is a time series problem for which a Long Short-term Memory (LSTM) neural network [2] is able to incorporate the time domain through recurrent connections. In order to decide whether the quality of water at time t is an anomaly or in a normal state, we assume that the relevant changes occur within the previous 30 minutes. The detector in this work consists of one LSTM layer and eight fully connected layers (10 hidden neurons each). The inputs to the LSTM network are all nine features from the challenge data set.

3 PRELIMINARY RESULTS

Figure 1, Figure 2 and Figure 3 show preliminary results for models trained with manual feature engineering and the automatic feature learning approach using a LSTM neural network. The models are trained on the data set provided by Thüringer Wasserversorgung for the GECCO IoT Challenge 2018 and tested using 10-fold cross-validation. The performance metric is the F1 score which trades precision and recall. The prediction of all classification models used in this work are of probabilistic nature. Hence, performance metrics are evaluated over the classification thresholds (i.e. a data point is classified as an anomaly when the prediction reaches the classification threshold).

4 DISCUSSION & OUTLOOK

The cross-validated F1 score of logistic regression with the original features in Figure 1 is below 10% for the standard 50% classification threshold and peaks at 25% for a threshold around 90%. The performance of logistic regression with the nine best engineered features in Figure 2 reaches 28% at a 50% classification threshold and even peaks at almost 50% at the 70% threshold. These results indicate that the features presented in this work indeed increase the performance of an anomaly detector. However, the LSTM neural network with automatic feature learning in Figure 3 has a stable classification behaviour with a peak F1-Score of 80% and shows a superior performance compared to the logistic regression models evaluated in this work. However, it is to be noted that logistic regression is used to determine if the engineered features increase the performance compared to the original features. The next steps are to evaluate the engineered features using non-linear models and support the feature space with anomaly scores from unsupervised detection models. A larger amount of samples through a greater time span is also expected to capture cyclic (e.g. seasonal) patterns.

REFERENCES

[1] Murray Gell-Mann and Seth Lloyd. [n. d.]. Information measures, effective complexity, and total information. Complexity 2, 1 ([n. d.]), 44–52. https://doi.org/10.1002/(SICI)1099-0526(199609/10)2:1<44::AID-CPLX10>3.0.CO;2-X

[2] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. Vol. 9. 1735–80 pages.