A self-consistent outlier filtering machine learning approach for predicting functioning of water pumps

Robbert-Jan Dikken

A self-consistent outlier filtering machine learning approach for predicting functioning of water pumps

Robbert-Jan Dikken*
Peutz bv, Computational Physics and Data Science, Zoetermeer, the Netherlands

(Dated: May 11, 2020)

The application of machine learning in predictive maintenance opens the way for significant cost reduction and reducing down time in industrial and infrastructural processes. Hidden relations in monitoring data can be utilized for fault prediction and optimization of maintenance strategies. However, in reality it often occurs that datasets are partially corrupted by faulty data. Because of the complexity that these datasets often have it is not trivial to distinguish incorrect data from correct data. In this article a dataset that represents the functioning of water pumps is analysed and a model is trained to predict functioning. A self-consistent outlier filtering approach is developed to handle incorrect data while training the model. This approach is validated by showing that the correlation between predictors and the water pump functioning parameter is significantly higher after the incorrect data is identified and removed from the dataset. The developed self-consistent outlier filtering approach can be applied in principle for any data driven modelling problem.

Keywords: machine learning, neural networks, data analysis, outlier filtering, maintenance

I. INTRODUCTION

Data science and machine learning are establishing a prominent role in our society. Applications range for instance from diagnosis and treatment plans in the medical field[1], optimization in the logistics sector[2], predicting properties of new materials in materials science[3] and optimization of building energy systems[4,5], to name only a few applications in a long list.

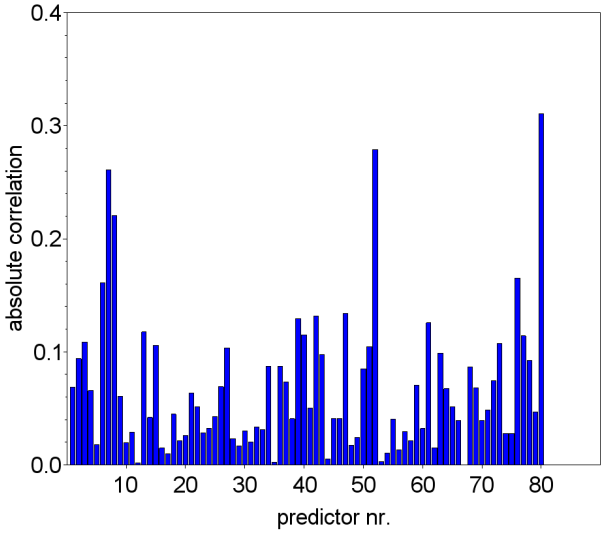
Also in infrastructure and the industrial and utilities sector there are ample possible applications of data science and machine learning. A good example is predictive maintenance[6–9]. Maintenance is conventionally based on statistics. However, more and more aspects of system operations are monitored. This opens the way to condition-based maintenance and predictive maintenance. This means that no longer the maintenance service has to rely on statistics which has a static character, but the maintenance strategy can be dynamic, based on the current or expected maintenance required, thereby saving costs and decreasing down time.

However, it is not at all a trivial task to assess based on monitoring data whether an installation or a component in an installation needs maintenance. This is due to the complexity that these datasets often have, but also due to inconsistencies in the data caused by mislabelling or malfunctioning of sensors. In this article a dataset that describes the functioning of water pumps is analysed and a model is developed to predict whether a water pump is functioning correctly, whether it needs repair, or whether it is not functioning at all. It will be seen that not all labelled data, through unknown cause, is consistent. The difficulty is to recognize which data is faulty and which data is correct. In this study the model itself will be used for this in a self-consistent approach. The resulting model can accurately predict the functioning of water pumps. Thereby, the method gives a framework for the ability to predict the functioning of water pumps with new data accurately.

II. DATA-ANALYSIS AND MODEL

The dataset includes predictors like geographic location, water quality, water quantity, management scheme, construction year, water pump type, etc. and the predicted feature being a label which can be 'functional', 'functional needing repair', or 'non-functional'. The dataset is considerably large with almost 60,000 samples of which 39% is labelled as 'functional', 7% is labelled as

FIG. 1. Absolute correlation of the predictors with the functioning parameter.



'functional needing repair' and 54% is labelled as 'non-functional'.

First the dataset is analysed by calculating the Pearson correlation coefficient of all parameters, shown in Fig. 1. It is found that none of the predictors has an absolute correlation with the water pump functioning parameter higher than 0.31, which means that there is no obvious predictor. Also there are no obvious correlations between the predictors. Therefore, no dimensionality reduction is performed on the dataset and the full dataset is used to develop a model to predict the functioning of the water pumps. To this end the predictors $\boldsymbol{x}$ are standardized,

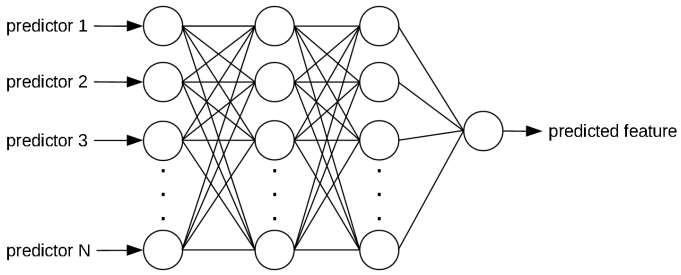
$$x_{\mathrm{i}}^{*} = \frac{x_{\mathrm{i}} - \mu_{\mathrm{i}}}{\sigma_{\mathrm{i}}}, \tag{1}$$

where μ_{i} and σ_{i} are the mean and standard deviation of predictor i, respectively, and the predicted feature y is normalized,

$$y^{*} = \frac{y - \min(y)}{\max(y) - \min(y)}. \tag{2}$$

The data is separated into a training dataset (80%) and a test dataset (20%). The training set is used to train the predictive model and the test set is used to independently validate the generality of the model. The model that is chosen for this specific classification problem is an artificial neural network that through supervised learning learns to classify the functioning of water pumps. A schematic representation of a neural network is given in Fig. 2. The activation function of the neurons in the network is given by a logistic function. A classical back-propagation algorithm[10] is applied on data batch sequences to update the weights between the neurons in the network. A small amount of noise is dynamically added to the training data to reduce the risk of overfitting.

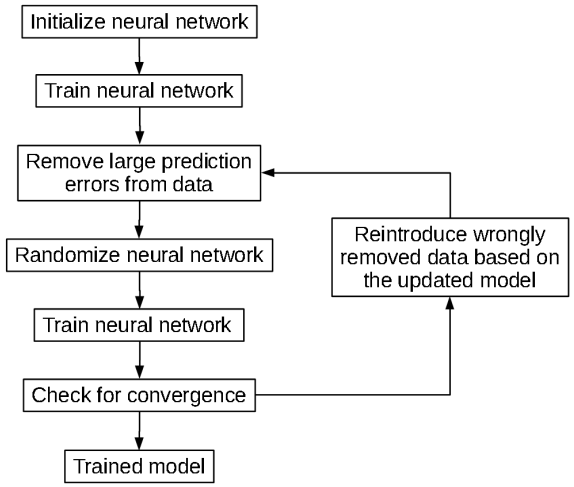
FIG. 2. Schematic representation of a neural network.



III. TRAINING A MODEL THROUGH A SELF-CONSISTENT OUTLIER FILTERING APPROACH

Because of the large number of variables that are involved in predicting the functioning of water pumps, it is extremely difficult to assess prior to training a neural network whether a part of the data is erroneous. The lack of clear correlation between predictors and target introduces the hypothesis that a non-negligible part of the dataset is faulty. Therefore we develop and apply a self-consistent approach in which we train a neural network first on the complete training dataset and assess its generality on a test dataset. Once the accuracy starts to converge, we use the output of the neural network to filter out the obvious outliers. Then the neural network is randomized again and retrained and assessed on the new training en test dataset, respectively. In each filtering episode it is checked whether in the last filtering episode data is wrongly removed based on the more advanced learning of the neural network. In this way the data and the neural network work together to clear the dataset of erroneous data. This self-consistent approach is schematically represented in Fig. 3.

FIG. 3. Schematic representation the self-consistent approach to outlier filtering.

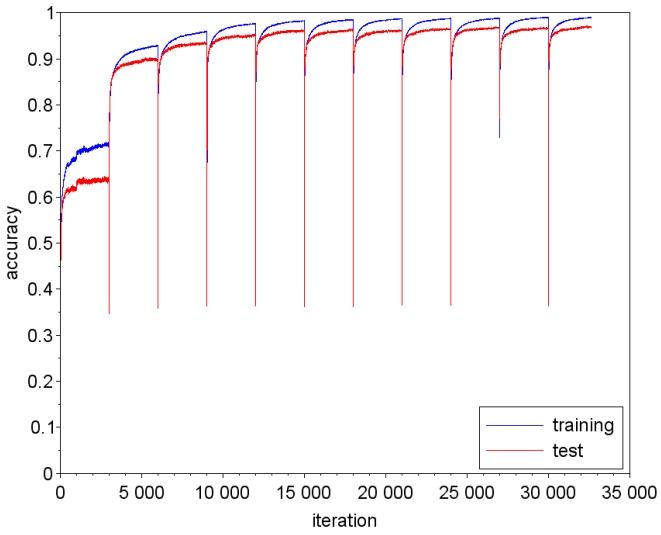


In Fig. 4 the evolution of the accuracy Q is given, defined as

$$Q = \frac{N_{\mathrm{corr}}}{N_{\mathrm{tot}}}, \tag{3}$$

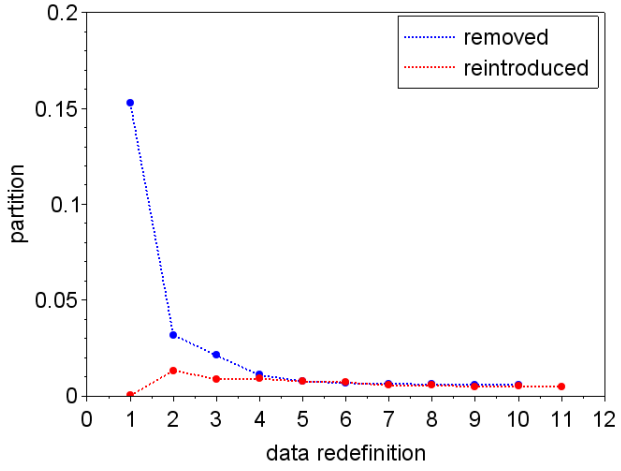
where N_{corr} is the number of correctly predicted samples and N_{tot} is the total number of samples in the dataset. The effect of data redefinition is mainly visible in the beginning of the curves. As the process of removing erroneous data and reintroducing wrongly removed data continues the effect becomes progressively smaller. The accuracy of the model reaches 0.99 and 0.97 for the training set and the test set, respectively. In Fig. 5 the removed partition and the reintroduced partition are shown as function of data redefinition. This explains the slowing convergence in Fig. 4. In the initial stages of the process it is clear what data should be removed and what data should be reintroduced. However, there develops an ambiguity at low partition value, which indicates that a low

FIG. 4. Evolution of the accuracy in both the training set and the test set.



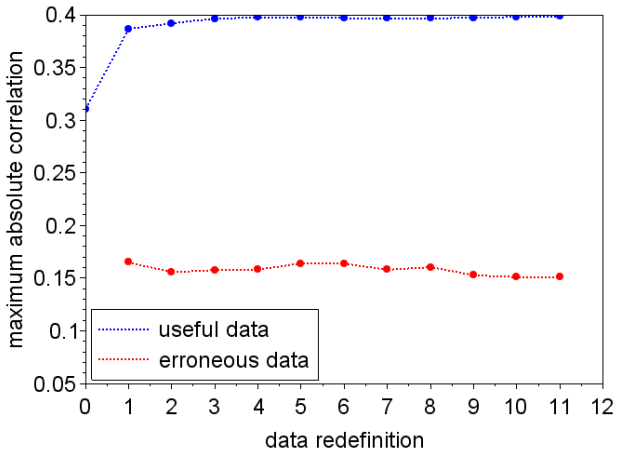
natural uncertainty of the model is approached, which coincides with the relatively high accuracy on both the training and test data set.

FIG. 5. Partition of the data being removed and being reintroduced through a self-consistent process.



To assess the validity of the self-consistent approach the absolute correlations of the predictors in the new dataset and in the removed dataset are computed. Figure 6 shows the maximum absolute correlation for both the new (useful) dataset and the removed data partition. The absolute correlation in the set of predictors with the water pump functioning parameter clearly increases and converges towards 0.4, while the absolute correlation for the removed data partition is considerably lower. This substantiates the thought that erroneous data in the dataset complicates the training of the model. The self-consistent approach in which the model itself is used to identify erroneous data is clearly effective.

FIG. 6. Maximum absolute correlation as function of self-consistent redefinition of the dataset showing both the determined useful partition and the erroneous partition of the data.



As a final test we tried to train a model specifically on the removed data partition. However, the best trained model on the removed data partition reached an accuracy of 0.68 for the training set, but only 0.45 for the test set. Also, training of the model on the removed data partition is prone to over-fitting, indicating a lack of generality in the removed data. These results are in agreement with the low correlation between predictors and the water pump functioning parameter. Hence, it is concluded that the self-consistent outlier filtering approach correctly identified and removed faulty data from the dataset, which makes the model more reliable.

IV. CONCLUSION AND DISCUSSION

In this study a method is introduced to handle erroneous data in large and complex datasets. Erroneous data complicates the training of predictive models and therefore it is desired to identify and remove these faulty data partitions. Due to the intricate nature of large datasets it is often not trivial to remove faulty data. Therefore, a method is developed that uses a self-consistent approach that filters outliers using a model that is simultaneously trained. Data that is found that is wrongly removed based on the continuously updating model is reintroduced. This sequence of checking for large prediction errors and checking for wrongly removal makes that both the model and the dataset evolve towards their optimum. The optimum for the model is such that it minimizes prediction error. The optimum for the model is such that it maximizes the correlation between predictors and predicted feature.

The method is applied to develop a model to predict, based on a large number of variables, whether a water pump is functional, whether it is functional but needing

of repair, or whether it is non-functional. Data analysis showed relatively low correlation between predictors and the water pump functioning feature. Through the self-consistent outlier filtering approach a partition of the dataset is found to be erroneous, for which it is shown that the correlation between predictors and the water pump functioning feature is significantly lower than in the final dataset. Such a model could be used for predictive maintenance of water pumps, which would greatly benefit maintenance cost and decrease down time.

The developed self-consistent outlier filtering approach can be applied in principle for any dataset or model. The condition that faulty data should not dominate the dataset, but is only a moderate partition, seems evident.

---

* r.dikken@peutz.nl

[1] M.K.K. Leung, A. Delong, B. Alipanahi, B.J. Frey, Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets, Proceedings of the IEEE, vol. 104, no. 1, pp. 176-197, Jan. 2016.

[2] S. Jimnez, T. De La Rosa, S. Fernndez, F. Fernndez, D. Borrajo, A review of machine learning for automated planning. The Knowledge Engineering Review, 27(4), 433-467, 2012.

[3] R. Ramprasad, R. Batra, G. Pilania et al., Machine learning in materials informatics: recent applications and prospects, npj Comput Mater 3, 54, 2017.

[4] R.J. Dikken, Introductie in machine learning en data science voor toepassing binnen de bouwfysica, Bouwfysica 3, 2018.

[5] R.J. Dikken, Machine learning en data science voor klimaatinstallaties, TVVL Magazine 7, 2019.

[6] Y. Peng, M. Dong, M.J. Zuo, Current status of machine prognostics in condition-based maintenance: a review. Int J Adv Manuf Technol 50, 297313, 2010.

[7] N. Sakib, T. Wuest, Challenges and Opportunities of Condition-based Predictive Maintenance: A Review, Procedia CIRP, Volume 78, 267-272, 2018.

[8] A. Van Horenbeek, L. Pintelon, A dynamic predictive maintenance policy for complex multi-component systems, Reliability Engineering and System Safety, Volume 120, 39-50, 2013.

[9] Vanraj, D. Goyal, A. Saini, S. S. Dhami and B. S. Pabla, Intelligent predictive maintenance of dynamic systems using condition monitoring and signal processing techniques A review, 2016 International Conference on Advances in Computing, Communication, and Automation (ICACCA) (Spring), Dehradun, pp. 1-6, 2016.

[10] T.M. Mitchell, Machine Learning, The McGraw-Hill Companies, Inc. (1997).