POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Machine Learning Data Driven Approach for Predictive Maintenance of Process Units

Master Thesis in Chemical Engineering

Author: Francesco de Fusco

Advisor: Prof. Flavio Manenti

Co-advisor: Andrea Galeazzi

Academic year: 2021-2022

1. Introduction

The new era of Big Data (BD) is driving the chemical industry in what has been called "The Fourth Industrial Revolution" [4]. In order to ensure a correct execution of the operations, chemical plants are monitored on a real time-basis and therefore produce a huge quantity of data daily. These latter are stored in a dedicated server and eventually processed for descriptive, diagnostic or prescriptive analysis carried out with statistical tools or dedicated software. The full exploitation of these data helps improving both profitability and productivity of the plant. In this perspective, chemical plants merge with digital technologies and BD giving life to Industry 4.0.

A well known problem in process industry is maintenance of the plant, responsible for huge economic losses especially in the petrochemical industry [5]. The implementation of BD technologies is causing a switch from traditional maintenance strategies to data driven approaches. This latter is based on statistical or Artificial Intelligence (AI) models able to overcome the problems resulting from mechanistic models obtained in laboratory environments and therefore barely applicable to real world scenarios. In recent years Gaussian Process (GP) have gained momentum in the world of Machine Learning (ML). Thanks to its Bayesian framework, this non-parametric model is able to identify patterns in complex time series data whilst prevent over-fitting. Gaussian Process Regression (GPR) has been successfully applied for industrial level predictions [2], whilst extensive research was done for a fully automation of the GPR [1].

Thesis objective

The aim of this thesis work is the development of an algorithm to be integrated in the Distributed Control System (DCS) of the Itelyum Regeneration plant located in Pieve Fissiraga (LO). The algorithm, developed in Python using the free, open-source library Scikit Learn, is able to collect the time series data from the Exaquantum Plant Information Management System (PIMS) installed at the plant; these are elaborated with ML techniques in order to return a predictive model descriptive of the current and future state of the process unit, thus allowing predictive maintenance. GPR model is compared with linear regression model in order to choose the best approach.

2. Methods

The modelling approaches of the time series data are based on Polynomial Regression (PR) and GPR. The models were trained with different ML techniques such as Cross Validation (CV) and ensemble method.

Linear regression

Given the simple mathematical formulation and the low computational power required for its learning, linear regression is used in many engineering and science applications. A particular case of linear regression, is the one in which the model is described by a polynomial of degree n plus the error ϵ

$$y = \theta_0 + \sum_{i=1}^{n} \theta_i x_1^i + \epsilon \tag{1}$$

In this case, the training phase consists in the estimation of the vector of parameters $\boldsymbol{\theta}$ by the well known minimization of the sum of squares. Once the regression is complete, it is possible extrapolate future values along with a confidence interval computed as

$$\bar{y_i^*} = \pm \delta \sigma_{err} \tag{2}$$

where σ_{err} is the standard deviation of the error and δ is a parameter that depends on the chosen confidence. The benefits of this model are its simplicity and negligible computational power required for computation. Conversely being a parametric model, it lacks of flexibility.

Gaussian Process Regression

Being non-parametric, the GPR does not rely on parametric assumptions, instead it adapts to the model complexity as more data arrive. Following a Bayesian approach, the GPR infers the function space by describing the distribution over functions. A GP is defined as

$$f(\mathbf{t}) \sim \mathcal{GP}(m(\mathbf{t}), k(\mathbf{t}, \mathbf{t}')) \tag{3}$$

where $m(\mathbf{t})$ and $k(\mathbf{t}, \mathbf{t}')$ are respectively the mean and covariance functions.
In practice given two data-points, a prior probability distribution is assigned to every function that could interpolate the points. Higher probabilities are given to functions that are considered to be more likely thanks to their properties. The combination of the data with the prior distribution leads to the posterior distribution that is the result of the regression.
In the case of GPR, the learning phase consists in finding the appropriate prior and its hyperparameters. The prior is defined by specifying the covariance function, also called kernel. This latter is chosen through a set of base kernels among which there are the Linear kernel (LIN), Rational Quadratic kernel (RQ) and Radial Basis Function (RBF) kernel. While, the LIN can model linear behaviours, RBF and RQ are able to detect wiggles and change of length-scales in the data. The base kernels are usually combined together by addition and multiplication to form a complex kernel whose characteristics are retained from the base kernels of which it is composed. The properties of the kernel are defined by its hyperparameters such as the length scale λ or variance σ^2. Along with the choice of the kernel, the learning phase implies the estimation of the vector of hyperparameters by minimization of the Log Marginal Likelihood (LML), defined as

$$log\left(p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta})\right) = -\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y} - \frac{1}{2}log|K_y| - \frac{n}{2}log(2\pi) \tag{4}$$

the disadvantage of using a GPR is the computational power required for the calculation, since it scales as $\mathcal{O}(n^3)$ [3].

Machine learning methods

In order to choose among the best model, this latter is trained on the train set, usually the 80% of the available data, and tested on the remaining 20%. The score on the test set is indicative of the goodness of the model. The evaluation metric used to compute the score is the Mean Absolute Error (MAE), that is the average of differences between predictions $\bar{y_i}$ and actual observations y_i:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \bar{y_i}| \tag{5}$$

This choice is justified by the presence of outliers and gross errors in the dataset. Indeed, the MAE does not penalize large errors too much.
In order to search for the optimal degree of the polynomial, an exhaustive search on all the possible combinations of the regressors up to a chosen order is done. Conversely, the domain of

possible combinations of kernel, creates a decision tree that is explored by a greedy search. That is, a path of optimum solutions is followed up to a desired depth of the tree. The deeper the tree, the more complex the kernel becomes. The greedy search is less effective but requires less computational power compared to exhaustive search.
In order to validate the models, Cross Validation (CV) is used. In CV, the training data is split into several folds and for each fold a training and test set is made. Since the dependency between the data in time series is high, the split must follow a temporal order and can not be done by random sampling. In CV, the final MAE is given by averaging the MAE computed on every fold. An ensemble forecast is a machine learning technique able to return models with higher robustness. The train set is sampled in different ways and different models are trained on the different samples. The final model will be a mean of the different models obtained.

Tools

The computational tool used for the development of the ML algorithm is the Python programming language, while Microsoft Excel was used as an interface between the DCS and Python. The Scikit Learn library was chosen for the implementation of the GPR and PR.

3. Dataset

The Itelyum Regeneration plant located in Pieve Fissiraga (LO) carries out the regeneration of exhausted lubricant oils to be reintegrated on the market. The process consists of three steps: pre-flash, thermal de-asphalting and hydrofinishing. Historical data of the plant were accessed thanks to the Exaquantum PIMS by Yokogawa that collects data form all the facets of the process and transform it in easily usable information. In addition to the data coming from the Exaquantum PIMS, further information on maintenance routines was integrated from an Excel sheet made available by the operators of the plant.
The unit taken under study is the furnace *PH-401B* in the thermal de-asphalting section, responsible for heating the dehydrated oil up to 360°C before distillation. More in details, the analyzed variables are shown in Figure 1. From above, it is reported a qualitative scheme

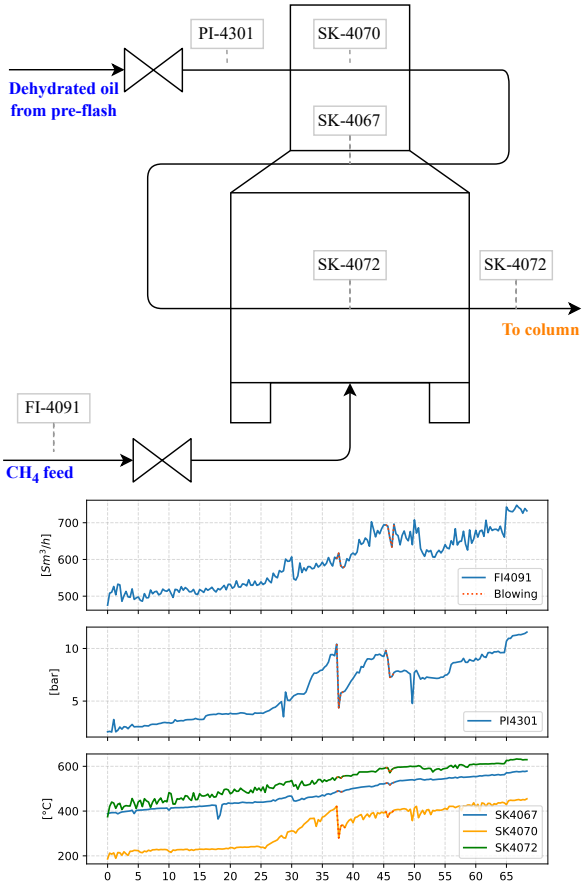

Figure 1: From above, qualitative scheme of furnace *PH-401B*, time series plot of CH_4 feed, pressure drop and tube skin temperature for lower, middle and upper section of the furnace. The blowing of tubes is highlighted in red, the hollows represent the downtime for maintenance.

of the furnace followed by the time series plot of methane feed *FI-4091*, pressure drop inside the tubes computed as $|PI4301\text{-}PI4304| \simeq PI\text{-}4301$ and skin temperature *SK-4072*, *SK-4070*, *SK-4067* of the tubes in the lower, upper and middle section respectively. The blowing, that is a maintenance operation during which the tubes are unclogged by coke deposition, is highlighted in red. The variable that is affected by blowing the most is the pressure drop. Right before the maintenance operation, the pressure reaches a peak and decreases sharply after the unclogging. On the x-axis it is reported the number of days after the startup of the furnace. Indeed from time to time, the furnace is stopped for general maintenance.

Experiment setup

In order to find the best approach to time series modelling, GPR and Polynomial Regression (PR) are trained with different ML methods as reported in the table below:

Model	Learning approach
PRCV	PR with CV
PRnCV	PR without CV
PRE	PR with ensemble and no CV
GPRCV	GPR with CV
GPRnCV	GPR without CV
GPRE	GPR with ensemble and no CV

Table 1: Description of the different learning approaches for training the PR and GPR models.

The best model is selected by simulating a daily refitting on the time series data of methane feed *FI-4091*. For every day, both GPR and PR are trained with the three different learning approaches to return a prediction on a forecasting horizon of size $h = 10 days$. The three approaches are than compared by means of a point wise MAE to monitor the daily performance and an average MAE to check the overall performance. This latter is computed as $\frac{1}{k}\sum_{n=1}^{k} MAE_i$, where k is the total number of days for which the models were refitted. Once the best learning approach is chosen, it is evaluated more in detail on the time series data of *FI-4091* and generalized on the pressure drop *PI-4301* and tube skin temperature of the middle section *SK-4067*.

4. Results

Learning approach comparison

The comparison of the three learning methods for the GPR is reported in Figure 2. It can be seen that GPRnCV has a lower average MAE with respect to the GPRCV and GPRE. Also, by looking at the second plot, it is possible to notice that the GPRnCV performs better the majority of times and its MAE is not fluctuating as for the other two models . Only in the last ten days, the GPRnCV shows higher MAEs values due to the presence of significant mean shifts in day 50 and 65 as can be seen from the first plot in Figure 1. For what concerns the train and test errors, the models seem to return a very good fit on the train set. This is caused by the high flexibility of the GPR that is able to model also the wiggles in the data. The results obtained for PR are similar and therefore not reported here. In conclusion, the best approach for the learning phase is with a simple train-test technique with a 80-20 split. CV and ensemble method seem to give worst results since this approaches use less training data for the learning step with respect to a simple train-test split.

Figure 2: Comparison of the performances of GPRCV, GPRnCV and GPRE. From above are reported the average MAE on train and test set and point wise MAE for every day of refitting. From the first plot, the GPRnCV seems the best choice. This is confirmed by the second plot since the GPRnCV performs better the majority of time. The difficulty in predicting the last 10 days is given by the irregular structure of the time series in that period.

Application of selected models

After simulating a daily refitting on the time series from *FI-4091*, GPRnCV and PRnCV were chosen as best candidate models for the prediction of the time series. Starting from above, Figure 3 shows the predictions returned by both GPRnCV and PRnCV on the three variable of interest *FI-4091, PI-4301* and *SK-4067* obtained 52, 39 and 53 days after the startup of the furnace respectively. The forecasting horizon is set to 10 days for *FI-4091* and *SK-4067* and to 5 days for *PI-4301*. Indeed, the time series of the pressure shows characteristics length-scales that are shorter than the one that characterize the other two variables. The flexibility of the GPR is clearly visible in the results obtained.

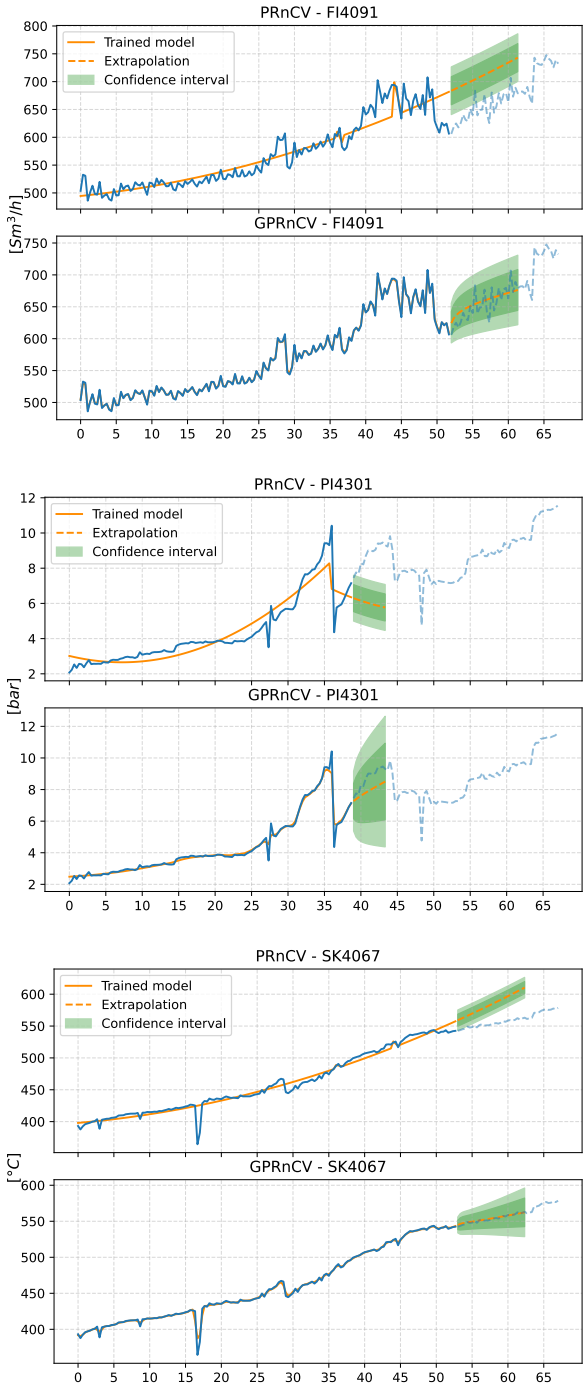

Figure 3: From above, comparison of predictions by GPRnCV and PRnCV on *FI-4091, PI-4301* and *SK-4067* obtained respectively 52, 39 and 53 days after the startup of the furnace reported on the x-axis. Being non-parametric the GPR is able to better fit the structure of the time series thus giving a better prediction. The prediction range is set to 10 days for *FI-4091* and *SK-4067* and 4 days for *PI-4301*. The shaded regions indicate 75% and 95% confidence intervals.

Thanks to the RBF and RQ kernels, the GPR is able to correctly learn the structure of the time series also when there are sudden changes in the trend such as for *FI-4091* and *PI-4301*. On the other hand, the PR regression can not adequately predict the value of the pressure even if the information about the blowing is inserted in the model as a categorical variable neither can model the sudden drop in the methane feed for which no assignable cause is available. Also by looking at the time series of the tube skin temperature, the GPR seems to be more responsive to the slight change in the structure of the time series.

Limit of the models

The limits of the models become clear when the prediction range becomes too large. To demonstrate it, an experiment involving a long term prediction was carried out. Both GPRnCV and PRnCV model are trained daily on the available data, while the prediction is carried out up to the day at which the furnace is turned off that is day 66. Therefore day by day, the available data increase while the prediction range decreases so, it is expected that the convergence of the model improves along time. Figure 6.6 shows the MAE obtained daily from the long term prediction on the variable of interest: FI-4091, PI-4301 and SK-4067. The dotted line indicates the value of the MAE to which the models converge, while the shaded region indicates the period for which the models do not converge at all to the final value. Both the GPRnCV and PRnCV model are able to predict the final state of the methane feed 15 days in advance with a mean average error of about 30bar. The high value of convergence is given by the sudden increase in methane feed right before the shutdown. For what concerns the pressure drop PI-4301, both GPRnCV and PRnCV converge to a low value of the MAE that is around $MAE = 1bar$. GPRnCV is able to return a good prediction 27 days in advance while PRnCV 19 days in advance. Also, the linear model seems to return a slightly better prediction. At last, the tube skin temperature before shutdown is adequately predicted 17 days before by the GPRnCV and 10 days before by the PRnCV with a value of the MAE lower than $10°C$.

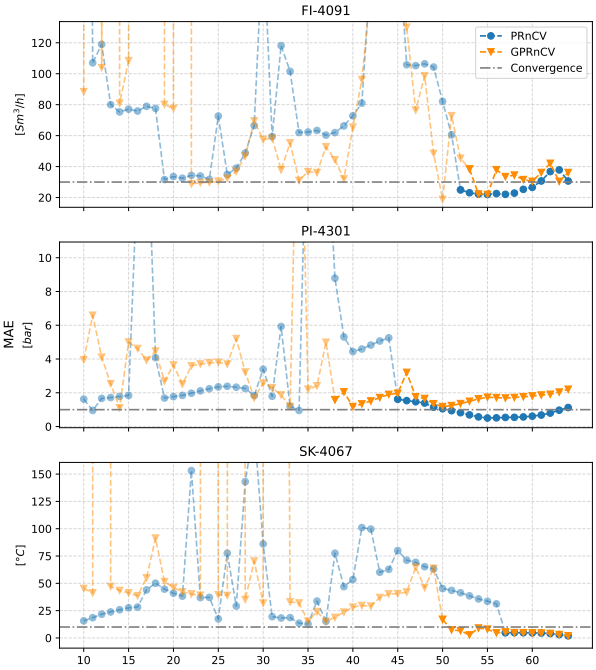

Figure 4: MAE resulting from daily long term predictions. The models are trained on the available data every day and extrapolate up to day 66 at which the furnace is shut down. On the x-axis, the days after the startup of the plant are reported. The dotted line indicates the value of the MAE to which the models converge, while the shaded region indicates the period for which the models do not converge.

5. Conclusion

This work provides a starting point to gain benefit from the massive amount of data generated over the years by the Itelyum Regeneration plant located in Pieve Fissiraga (LO). The study aimed at developing a data driven approach to predictive maintenance of the process furnace located in the thermal de-asphalting section of the plant. This was done thanks to time series modelling of methane feed, pressure drop and tube skin temperatures in order to extrapolate the future state of the variables. By comparing PR and GPR, it was found out that the latter is most suitable for modelling the time series coming from the process furnace since they show high irregularities and no well defined structure that is captured by a non-parametric and flexible model as the GPR. On the other hand, the PR model can only capture these irregularities by passing a categorical variable that can be difficult to obtain and not always accurate. Also, the approach used to train the GPR, that is a greedy search tree, allows for a good generalization of the model on time series different from that on which the model was trained. In conclusion between GPR and PR, the former is the best candidate to develop a data driven approach modelS to predictive maintenance

References

[1] David Duvenaud. *Automatic Model Construction with Gaussian Processes*. Thesis, University of Cambridge, November 2014.

[2] Zhiqiang Ge, Tao Chen, and Zhihuan Song. Quality prediction for polypropylene production process based on CLGPR model. *Control Engineering Practice*, 19(5):423–432, May 2011.

[3] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass., 3. print edition, 2008.

[4] Klaus Schwab. The Fourth Industrial Revolution: What it means, how to respond. page 7, January 2016.

[5] Venkat Venkatasubramanian, Raghunathan Rengaswamy, Kewen Yin, and Surya N. Kavuri. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers & Chemical Engineering*, 27(3):293–311, March 2003.