Smart Alarm: alarm prediction using sensor data

Ahamed Buhari and Chandraprakash Sahu
Supervisors: Hadi Fanaee and Mahmoud Rahat

1. Problem Formulation

We will work with a sub-sample of data retrieved from Alfa Laval's IoT portal, which corresponds to data of 27 water purifier machines installed on different ships. These machines purifie oil and water supplies onboard marine vessels. In total we have 102 sensors, and 115 parameters that are set by user can be modified at different situations. The data is collected for a period of 3-8 months, depending on the machine.

The company is interested in developing a smart alarm system that predicts alarms and warnings in advance before their occurrence based on sensor readings. So, the main objective of our project is to predict alarms and warnings using machine learning techniques and find the best predictive models. The final goal is to exploit these models on the machine's controller to make real-time predictions. Presently, these alarms are functioning based on predefined threshold limit and once it crosses the limit, the system generates fault and warning messages. Predictive models will improve the useful life of the machines, avoids downtime, enhance the product quality and quantities, and prevents damages or faults.

More specifically, we deal with the following research questions:

RQ1: What is the ideal window size for taking samples before occurrence of alarms and warnings?

RQ2: What sampling frequency (e.g. 5 seconds, 1 minutes, 5 minutes, etc) in time series is ideal for making better predictions?

RQ3: Which sensors and parameters contribute more in making better predictions?

RQ4: Which machine learning method provides better predictions (e.g. classical or sequential)?

RQ5: Which set of configurations makes a better predictive model?

We will investigate different machine learning techniques and find which gives the better model. In particular, we would like to compare classification algorithms with deep learning-based models such as RNN and CNN and also classical time series classification methods such as

HMM. We will use standard machine learning performance metrics such as accuracy, precision, recall, F-measure, confusion matrix and ROC curve for comparison of the built models.

Novelty

This is an experimental research project with high industrial impact. So far company has not used any predictive models on their data. So, making accurate predictions can provide added value to machines and provides profits for the company and the customers.

Unfortunately, currently there is no off-the-shelf machine software for doing direct machine learning on this type of data (multi-sensor unequal-length time series), so one main part of the challenge is feature engineering and extraction and make the data prepared for machine learning models.

2. Literature Review

From the application point of view, it is related to an area of research called predictive maintenance [1]. From methodological point of view this thesis is about (multivariate) time series or sequence classification [2,4], and (Semi-) supervised anomaly detection [3]. So, the related work in the broad term includes any work related to these areas.

Anomaly detection deals with identification of events that gives rise to the uncertainties due to a substantial deviation. The methods for anomaly detection can be classified into supervised and unsupervised approaches. In supervised methods the model is fed with labeled data of both normal and abnormal instances and then model can be used for predicting abnormal instances. On the contrary, unsupervised methods do not use labelled data and they can identify anomalies by just comparing them with the major part of the data which is assumed to be normal.

One of most naïve solution in unsupervised anomaly detection for conditioning monitoring is to use statistical process control techniques on the sensor measurements along with a threshold limit. For instance, in [5] authors used this method for anomaly detection in manufacturing process.

From supervised category typically machine learning is trained with normal examples and then the model is used on the test examples to identify those examples that deviates from the major class. For instance, researchers in [6] use a fully convoluted time series classification. The original time series is split into several equal length sub-sequences. To construct a new sequence, a statistical feature on each subsequence is extracted to evaluate the classification efficiency of test sets, sample sequence and precision of network classification failure with a fully convoluted network using the original sequences.

Another popular method for time series classification are sequential models. For instance, in [7] a sequential model is used to make predictive models from time series.

3. Project Plan

There are few fundamental key steps have been planned to achieve the goal of this project, which is shown in the given figure as well.

2.1 Project Understanding: The first step is to gather the information of the project which helps to design blueprint of the project.

2.1 Data Understanding: Getting familiar with data is one of the important sections, it gives first insight knowledge about the data quality, features information and its connection with other subsets and directions to each target feature.

2.3 Data Preparation: This is the most challenging and crucial part of the project. If the data is prepared correctly, It will give the best result as desired output else it will give poor results.

2.4 Data Validation/modelling and Evaluation: In this step, splitting the data in training/testing and Validation. After training the model by performing various machine learning algorithm in the training data set, validating the result in validation dataset until reaching the suitable output. After that, testing the model in test dataset to predict and evaluate the output.

The final evaluation of the project is done by comparing the best accurate model which gives better performance in the test data set.

Project Planning is a crucial part of accomplishing the project. It is a baseline of a project which gives complete information about the project strategy. In methodology following activities have been planned to complete the project task with respect to time plan.

The following activities have been listed for time for achieving the goal in the given period, including the three milestones as a project presentation are planned which tracks the project ongoing status, it describes as follows. First, at the beginning of the project, it is an initial level of the project presentation which describe the project, team and goal information of the project thesis.

The second presentation is planned in the mid of the project in March, which describes the ongoing status flow of the project, challenges faced during the project and final projection of the thesis.

The final presentation is planned for the last week of May, which describe related to final project implementation and its results

	Complete thesis writing and Submission
M	Thesis Presentation
A	Modification/Correction
Y	Review and Feedback from Supervisor
A	Final Thesis Writing
P	Model Improvement
R	Feedback from Supervisor
M	Model testing and Evaluation
A	Half Time Seminar
R	Model Formulation
F	Feedback from Supervisor
E	Model Novelty Evaluation
B	Novelty Preposition
	Implementation of Algorithm
	Modification/Correction
J	Feedback from Supervisor
A	Finalization of Algorithm to Use
N	Learning Method Shortlist
D	Data Preprocessing Task
E	Dataset Observation
C	Initial Level Thesis Report Submission and Presentation
	Literature Review
N	Project Introduction
O	Project Approved
v	Project Selection

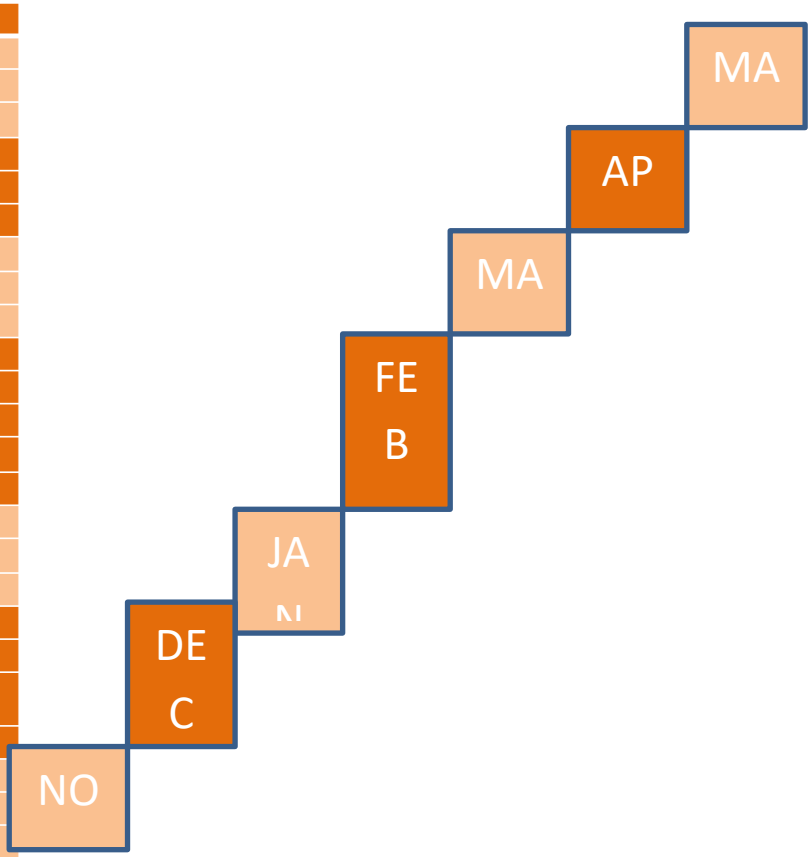

Figure 7. Project plan

In this project thesis, the Agile methodology is used for delivering the project incrementally within time. It is one of the most preferable because it has several benefits compared to other methodology used. Fast processing, accelerating decision making, the level of communication within the project, attention to technical excellence and best design in the project flow. It divides the complex project in short sprints which aids to achieve the goal effectively. It keeps updating the ongoing project activities to all the team members and many more advantages to other methodologies. the process flow diagram and steps to get an effective model in the project.

.

4. Bibliography

[1] Zhang, Weiting, Dong Yang, and Hongchao Wang. "Data-driven methods for predictive maintenance of industrial equipment: A survey." IEEE Systems Journal 13.3 (2019): 2213-2227.

[2] Xing, Zhengzheng, Jian Pei, and Eamonn Keogh. "A brief survey on sequence classification." ACM Sigkdd Explorations Newsletter 12.1 (2010): 40-48.

[3] Cook, Andrew, Göksel Mısırlı, and Zhong Fan. "Anomaly detection for IoT time-series data: A survey." IEEE Internet of Things Journal (2019).

[4] Wei, Li, and Eamonn Keogh. "Semi-supervised time series classification." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006.

[5] Kamat, Pooja & Sugandhi, Rekha. (2020). Anomaly Detection for Predictive Maintenance in Industry 4.0-A survey. E3S Web of Conferences. 170. 10.1051/e3sconf/202017002007

[6] Lei, Yuxia, and Zhongqiang Wu. "Time series classification based on statistical features." EURASIP Journal on Wireless Communications and Networking 2020.1 (2020): 1-13.

[7] A. Horelu, C. Leordeanu, E. Apostol, D. Huru, M. Mocanu and V. Cristea, "Forecasting Techniques for Time Series from Sensor Data," 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, 2015, pp. 261-264, doi: 10.1109/SYNASC.2015.49.