Industrial Assets Performance Labelling Based on Numerically Encoded Event Logs

Pierre Dagnely[1], Tom Tourwé[1], and Elena Tsiporkova[1]

Sirris - Elucidata Innovation Lab, A. Reyerslaan 80, 1030 Brussels, Belgium,
(pierre.dagnely, tom.tourwe, elena.tsiporkova)@sirris.be

Abstract. Assessing the performance of industrial assets usually requires exploring and combining sensor data, event logs, asset characteristics and domain expert knowledge. Therefore, this process is time and resource consuming. Extrapolating the performances solely from the event logs could lead to more optimal/pro-active planning of maintenance activities. In [1], we have shown that event logs could be numerically encoded into event profiles accurately representing asset event behavior. Therefore, it is possible to extract the event profile of a new operational cycle and link it with the similar event profiles of past operational cycles for which the performance is known. It offers a gain of time and resources when exploring the performance of new operational cycles. We propose a methodology to label asset performances solely based on the event logs, using a standard (numerical) classifiers. The performance of a new asset operational cycle can then be assessed with negligible computational time. The methodology is validated on real-life data from a photovoltaic plant.

1 Introduction

Fast labelling of the performance of new industrial runs, i.e. operational cycles is important for monitoring and maintenance purposes. If the performance of an asset can be quickly assessed, maintenance activities can be planned responsively, and the asset unavailability will be minimized. Moreover, if the performance can be labelled with a profile indicating the root cause, then the maintenance team will already have insights on the problem and its potential solution before arriving on site. It would lead to more optimal planning of maintenance activities.

However, extracting performance profiles indicating root causes is a complex and time-consuming task. For instance, we have shown in [1], that in the photovoltaic (PV) domain, it requires to combine irradiation data, yield sensor data, event logs, plant characteristics and domain expert knowledge. Such process is time and resource consuming and can not be done every night to assess the performance of each of the thousands of plants in a portfolio.

Therefore, methods that could quickly label asset performances would be very valuable in industrial domains. One solution is to rely on the valuable information provided by the event logs. Meaningful event profiles can be extracted from the event logs. For instance, one profile would be mainly characterized by the occurrences of the event "over-temperature" while the other would be characterized by the occurrences of the events "under-temperature" and "sensor test". These event profiles represent various internal behaviors. Then they can be labelled using sensor data and domain knowledge. For instance the over-temperature profile could be characterized as under-performing while the other would be characterized as a regular behavior. Subsequently, they can be used to characterize the performance of new operational cycles. A classifier can evaluate if the event profile of a new operational cycle is similar to one of the past event profiles for which the performance has already been evaluated. In this way, the performance of the new operational cycle of an asset can simply be derived from its event log through a standard classifier and without the need of inputs from domain experts.

Such approach has been successfully explored by Fronza et al. [2] who have generated numerical event profiles using random indexing (RI). They have trained an SVM classifier to assess the performance of software runs as faulty or not, solely based on their event logs. Unfortunately, the performance of their method on our validation dataset has been disappointing since it relies on training the models on software runs ending in failures (stops) which can be potentially predicted by the preceding contextual information. However, this is not always the case in many industrial domains where assets produce a lot of warning and errors, but failures occur very rarely. Therefore, we propose in this paper an alternative approach more suitable in the given context.

The paper is organised as follows. First, the relevant literature about text classification is explained in Sections 2. Then, in Section 3, we explain our methodology for fast labelling of new operational cycles. In Section 4, we validate our methodology in the PV domain by developing a case study on a real-world dataset provided by our industrial partner. Finally, we conclude the paper with a discussion and an outlook for further research in Section 5.

2 Literature Review

One of the challenges when applying classification to event logs is their textual nature while most classifiers have been built to handle numerical values. The typical approach in text classification is then to extract numerical features from the text and apply the classification on these features. For instance, the features could be the total number of words in the documents, the average length of the words used in the documents or the total number of punctuation marks in the documents. However, these methodologies can be complex to deploy. As stated by Dalal et al. [3], they would require to be tailored to the event behavior, e.g. "do all events have the same impact?" or "is the repetitions of the same event relevant?", and would likely have to include meta-data. Therefore, textual classifiers pose a challenge to develop an agnostic methodology. On the other hand, numerical classification methodologies are well defined, more agnostic and validated in various domains. These methods can be applied to various numerical data, if the training dataset has been well constructed.

Therefore, another approach is to numerically encode the event logs. Fronza et al. [2] have applied such method using RI to numerically encode the events logs. RI is a data reduction method from the text mining field proposed by Sahlgren in [6]. This method is used to store in a condensed way the “context” of a word, i.e. the surrounding words. Fronza et al. have applied this methodology on event logs generated by software by considering each event as a word and each event log as a text. They were able to classify the software runs as faulty or not faulty with a high accuracy.

3 Methodology

We have developed a methodology to label performance of asset operational cycles based on their events logs. The steps of our methodology are the following:

1. Convert operational cycles to numerical standardized profiles, i.e. relevancy score vectors
2. Annotate the relevancy score vectors with performance labels
3. Train a classifier to label relevancy score vectors
4. Extract relevancy score vectors of new operational cycles and label them on the fly using the classifier

3.1 Relevance Score Extraction

The main challenge faced by our methodology is the textual nature of the event logs which hinders their processing. Extracting typical event logs, i.e. event profiles, from textual event logs would require domain knowledge. However, we intend to minimize the need of domain expert inputs as their time is valuable. We solved this problem by numerically encoding the event logs as relevancy score vectors. It allows to build agnostic clustering from the event logs. The relevancy score methodology that we defined in [1] is applied on the event logs. Our methodology follows 2 steps: 1) The event logs are segmented by operational cycles; 2) The relevancy scores are computed based on the event frequency.

Defining atomic event logs The first step is to divide the event logs into atomic pieces, i.e. into "traces or meaningful periods" of the asset, called atomic event logs (AEL). For instance, in case of a car, the event logs could be divided into operational cycles, from the start of the travel to its end. The definition of these atomic event logs is therefore domain and goal oriented. The main interest is to transform the continuous stream of events into a meaningful finite set of event logs. These AELs will be easier to analyze and interpret. In addition, they contain all the event correlations. For example, the interpretation of the event

”temperature error” is modified in case it is preceded by the event ”temperature sensor broken”. The goal of the segmentation into AELs is to have the events that could interact all stored together in one file.

Computing relevancy score We have used a method inspired by the widely used in text mining TF-IDF score, where for each event type of each AEL, its relevancy score is computed. The goal is to attribute a score reflecting the "abnormality" of the event, i.e. determine degree to which the event deviates from the regular asset behavior. For example, the critical event "temperature error" that occurred 2 times in the atomic logs should have a high relevancy score as it indicates a failure, while the event "start" (representing the usual behavior of the device) that occurred 17 times should have a relevancy score of 0. Therefore, the events' frequencies need to be carefully exploited.

By considering the AELs as a text, text mining methods such as TF-IDF can be adapted for this purpose. Therefore, our methodology relies on the computation of two frequencies: 1) The frequency of the event (type) in the AEL, and 2) the frequency of the event (type) in well selected corpus of AELs aligned with the analysis goal in mind.

First, the term frequency (TF) is computed, i.e. for each event type that can be reported by the asset, its frequency in the AEL is computed. The formula below is used.

$$TF_{e_i,a_i,l_i} = \frac{\#\text{ occ. of events } e_i \text{ in logs of asset } a_i \text{ for AEL } l_i}{\#\text{ of event in AEL } l_i \text{ for asset } a_i}$$

The inverse document frequency (IDF) need to be adapted to the industrial event logs context as the text corpus on which it relies does not apply here. Therefore, the corpus definition needs to be adapted. Three approaches are possible and need to be carefully selected:

- The corpus consists of all available AELs. It allows to compare asset behavior over time and across assets.

$$IDF_{e_i} = \log \frac{\#\text{ of AEL in all assets and all days}}{\#\text{ of AEL where event } e_i \text{ occured}}$$

- The corpus consists of all the AELs of one asset. It allows to focus on one asset behavior and monitor the evolution of performance over time.

$$IDF_{e_i,a_i} = \log \frac{\#\text{ of AEL for asset } a_i}{\#\text{ of AEL for asset } a_i \text{ with event } e_i}$$

- The corpus is composed of AELs of all assets for the same trace (e.g. the same day). It allows objective comparison of performance across assets. However, as events occurring in all AELs of the corpus are considered less relevant, a failure occurring in all assets would be masked by this case.

$$IDF_{e_i,p_i} = \log \frac{\#\text{ of AEL occuring at the period } p_i}{\#\text{ of AEL for period } p_i \text{ with event } e_i}$$

Subsequently, the relevancy score is computed by multiplying TF and IDF: Relevancy score $= TF_{e_i,a_i,l_i} * IDF$

In this way, the relevancy score uses the frequency of the event (more frequent events have higher scores) corrected by the IDF that will decrease the score of events frequent in the corpus (if an even occurs in all AELs of the corpus, its IDF is $\log(1) = 0$, which leads to a relevancy score of zero).

By computing the relevancy scores over all events for each operational cycle, a numerical vector representing the asset events relevancy, i.e. event profile, for the operational cycle is obtained. The textual representation of the events has been transformed into a numerical feature vector and the event logs have been preprocessed by assigning null (or low) scores to the less relevant events.

The advantage is twofold: 1) the numerically encoded event logs can easily be clustered using traditional clustering methods. 2) event logs of various sizes are converted into numerical vectors of fixed size, which facilitates their comparisons.

3.2 Performance Annotation

One of the main challenges in industrial setting is the lack of labelled data. Therefore labelling strategies needs to be applied e.g. a rather straightforward approach is to label the past AELs as faulty, i.e. if an outage occurred, or healthy, i.e. if no outage occurred, based on domain experts inputs. More advanced labelling can also be performed, combining multiple data sources. In [1], we labelled PV data by combining sensor data, event logs, plant characteristics and domain expert knowledge. This time and resource consuming methodology allowed to label historical data with complete and precise performance profiles such as “Inverter-days with high outages due to Riso Low, mainly occurring in the end of the summer”. Time consuming methods providing detailed labels can be afforded in this step as the labelling only need to be performed once on the historical data.

3.3 Classifier Training

The goal is to build a model classifying the relevancy score vectors into their performance profile, e.g. classify the vector, i.e. AEL, as healthy or faulty. Relevancy score vectors being numeric, they can easily be classified by a numerical classifier. Moreover, the relevancy score methodology not only converts the event logs as numerical inputs, but also pre-processes the event logs. It hides the irrelevant events and pinpoints the most important ones. Therefore, the combination of classification methodologies and relevancy score allows an easy classification of textual event logs. Numerical, widespread and agnostic classifiers can be applied to the event logs to build classification models.

3.4 Labelling new operational cycles

The model can then classify a new AEL performance, solely based on the event logs. The relevancy score vector of a new AEL can be extracted from its event log. The classifier can then detect to which performance label it belongs and hence what is the performance of the new AEL.

4 Evaluation and Discussion

The critical performance aspect of our methodology is the ability of the classification algorithm to correctly label the new incoming operational cycles. We have experimented with several different widely used classifiers and compared their accuracy. Another factor impacting the overall performance of our methodology is the encoding of the textual event logs into numeric relevance vectors based on TF-IDF scoring. We consider important to benchmark our methodology with the one developed by Fronza et al. [2] using an alternative encoding approach based on RI.

4.1 Data understanding

We have used one year of event logs from one - often faulty - PV plant. The data has been provided by our industrial partner 3E, which is active, through its Software-as-a-Service SynaptiQ, in the PV plant monitoring domain. PV plants are composed of several PV modules (that convert the irradiation into direct current) connected to one or several inverter(s) (that convert the direct current to alternative current) which send the current to the grid. These systems are now continuously monitored. In addition, various sensors (measuring the irradiation, electricity production, ...) are present in the plant. An inverter reports status, i.e. its current state like start, stop or running, but also other events that can represent e.g. an outage (such as grid fail or string disconnected) or other phenomena (such as over-temperature or DC current under threshold).

In the PV case, an AEL corresponds to the event logs of one inverter for one day. As the plant is only active during the day, it can be considered that it "reboots" at night (often, small problems disappear the next morning).Therefore, each day corresponds to a operational cycle. In addition, as the events are monitored and reported at the inverter level, AELs are at the inverter level. Hence, for our plant with 26 inverters, 9490 AELs have been obtained (26 inverters-logs times 365 days). An AEL typically contains around 5 distinct event types. Over our one year dataset, 54 distinct event types have been reported. Therefore, our relevancy score vectors have a length of 54 but only around 5 non null values.

Each AEL has been labelled as healthy or faulty using additional datasets. An AEL is labelled as faulty if an outage is detected during that operational cycle, i.e. if there was enough irradiation to have electricity production but there was no yield (for at least 30 minutes). There are 3237 healthy AELs and 539 faulty AELs. Note that it creates an unbalanced dataset with 85% of healthy AELs. Therefore, a classifier labelling all AELs as healthy would be correct 85% of the time. Note that events related to a failure only occurred in the AEL where the failure occurred, allowing to randomly split our dataset into training and testing datasets and apply 10-fold validation.

4.2 Classifier Performance Benchmarking

We have benchmarked 7 classifiers namely: 1) Logistic regression, 2) Linear discriminant analysis, 3) K Nearest Neighbor (kNN), 4) Decision Tree (DT), 5) Random Forest (RF), 6) Gaussian Naive Bayes and 7) Support Vector Machine (SVM). These methods have been selected as they are widespread and have been validated in various domains [7, 8]

Figure 1 shows the mean accuracy (in abscissa) for each classifier (in ordinate). The accuracy is computed using a 10-fold validation methodology, i.e. the dataset is randomly divided in two datasets, 90% used as training and 10% used as testing and the accuracy of each classifier is assessed on this dataset. Then the process is repeated ten times to ensure statistical significance of the accuracy. The accuracy is computed using the Jaccard similarity score traditionally used for binary classifier. The score computes the overlap between the prediction of the classifier and the reality. A score of 1 represents a perfect classification, i.e. a 100% overlap between prediction and reality. The computation time of the methods is not considered in this evaluation as it was negligible on our dataset.

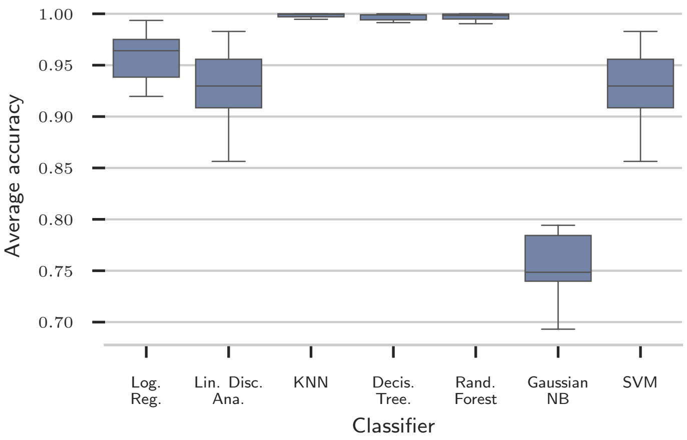

Fig. 1. Accuracy of the 7 classifiers applied on the event logs pre-processed through the relevancy score methodology

Figure 1 indicates that the classifiers kNN, DT and RF are the most accurate, with an almost perfect accuracy of around 0.98. Gaussian NB has the lowest accuracy with a mean accuracy of 0.75. SVM, Linear discriminant analysis and Logistic regression have similar accuracies, with a mean accuracy of around 0.94. From this figure, it appears that DT, RF and kNN are probably the best methods to consider for our workflow.

Table 1 contains the precision (the ability of the classifier not to label as positive a sample that is negative), recall (the ability of the classifier to find all the positive samples) and F1 score (the combination of precision and recall) for the three pre-selected methods. It appears that kNN and DT are slightly superior to RF in terms of recall for the faulty AELs. RF only retrieves 0.72% of the faulty AELs while kNN retrieves 0.96% of them and DT 95%. kNN and DT have otherwise similar accuracy with kNN slightly more accurate. The remaining metrics are similar. However, as it is important to not miss any faulty AEL, the recall score for the faulty class is the most important metric. Therefore, kNN (using 3 neighbors) is the most suited classifier.

This validation showcases the accuracy of our methodology to recognize faulty or healthy event logs. It indicates that our relevancy score vectors are able to hide the irrelevant events and pinpoint the relevant ones, i.e. the ones related to healthy or faulty behaviors. The classifier can then easily label new asset AELs based on these few relevant events.

Table 1. Comparison of the accuracies for Random Forest, Decision Tree and K Nearest Neighbor

Classifier	Faulty class			Healthy class		
	Prec.	Recall	F1 score	Prec.	Recall	F1 score
Random Forest	**1.00**	0.72	0.84	0.96	**1.00**	0.98
Decision Tree	0.96	0.95	0.95	**0.99**	0.99	**0.99**
K Nearest Neighbor	0.98	**0.96**	**0.97**	**0.99**	**1.00**	**0.99**

4.3 Relevancy Score vs. Random Indexing Encoding

A similar methodology has been successfully applied by Fronza et al. [2] on software event logs using RI. Benchmarking our methodology against the methodology of Fronza et al. would allow to compare the accuracy of the relevancy score methods against RI. Subsequently, any difference is supposed to come from the numerical encoding method or the type of data (software vs. industrial (PV) asset), as the methodologies are otherwise similar.

We have computed the mean accuracy of the methodology of Fronza et al. for the 7 classifiers experimented previously. They have been applied on the same dataset with the same 10-fold validation methodology.

Figure 2 shows the results for the 7 classifiers. It shows the mean accuracy (in abscissa) for each classifier (in ordinate). Except for Gaussian NB with a mean around 0.1, the other methods have a similar accuracy of around 0.50 - 0.60. The RI approach seems therefore less accurate than the relevancy score one, for each classifier. In their study, Fronza et al. obtained a mean accuracy of 0.80-0.90, so a better accuracy than in our dataset. It indicates that we have a different event behavior in the PV domain which disadvantages RI.

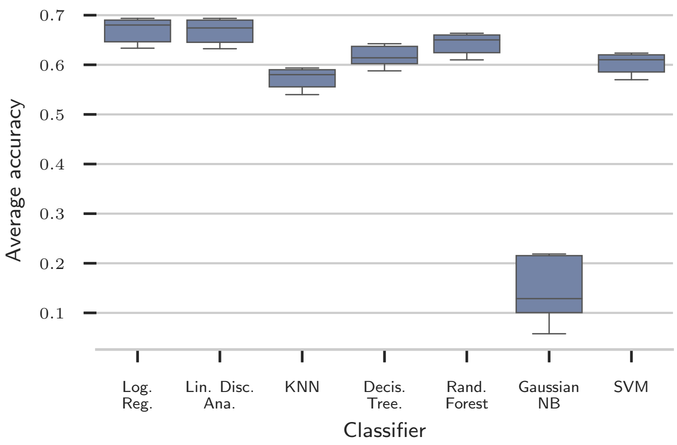

Fig. 2. Accuracy of the 7 classifiers applied on the event logs pre-processed through the RI methodology

The cause of the low accuracy of RI is twofold. First, RI is less adapted to event behaviour. Software are procedural, i.e. events typically follow the same sequence when an action occurs. On the other hand, inverters (and many other industrial assets) may report events in various order for the same action. In addition, software events have strong connections with the previous events reported, sometimes with events reported long time ago. PV events usually only interact with a few events in their immediate neighbourhood. Our methodology does not focus on event correlation but on pinpointing the relevant events and is therefore not impacted by the variation in the event order.

Second, Fronza et al. defined an operational cycle/AEL as the set of events generated by the software from its start until its end (labelled as healthy) or until a failure (labelled as faulty). Therefore, faulty software AELs always end with a failure. However, the labelling of industrial asset performance is less strict or precise. In our dataset, healthy and faulty periods can co-occur in a faulty AEL. RI seems therefore less resilient to less precise labelling of the data as it considers all the events.

Therefore, RI is more suited for procedural data with strong context-awareness while the relevancy score methodology is more adapted for variable noisy data, i.e. data with many irrelevant events.

5 Conclusion

Our methodology allows rapid annotation of new asset operational cycles with a known performance label or profile, only based on event logs. Assessing the performance of an asset is time and resource consuming as it implies analyzing the event logs, the various sensor data, the asset characteristics and requires domain experts' knowledge. We have shown that our methodology was able to label performance of new operational cycles with a mean accuracy of 98% using a kNN classifier, solely based on the event logs and with negligible computation time. Moreover, our methodology has been shown as more effective than the Fronza methodology, relying on RI, for our validation domain. It ensures a fast and scalable labelling of the asset performance that could be deployed in industry.

In terms of further research possibilities, our methodology should be validated on other datasets to benchmark RI and relevancy score methodologies. Both methodologies seem to have different accuracies based on the application domain. We suspect that RI is more suited for procedural data with strong context-awareness while the relevancy score methodology is more adapted for variable noisy data, i.e. data with many irrelevant events. A more thorough comparison of the two approaches on event logs generated by various asset types e.g. cars, software, medical records, ... could allow to assess this assumption and better estimate the event characteristics that impact the accuracy of both encoding approaches. Moreover, it could allow to further refine the classifiers accuracies and better understand them.

Acknowledgements

This work was subsidised by the Region of Bruxelles-Capitale - Innoviris.

References

1. P. Dagnely, T. Tourwé, and E. Tsiporkova, "Annotating the Performance of Industrial Assets via Relevancy Estimation of Event Logs," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 1261–1268.
2. I. Fronza, A. Sillitti, G. Succi, M. Terho, and J. Vlasenko, "Failure prediction based on log files using Random Indexing and Support Vector Machines," *Journal of Systems and Software*, vol. 86, no. 1, pp. 2–11, Jan. 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0164121212001732
3. M. K. Dalal and M. A. Zaveri, "Automatic Text Classification: A Technical Review," *International Journal of Computer Applications*, vol. 28, no. 2, pp. 37–40, Aug. 2011. [Online]. Available: http://www.ijcaonline.org/volume28/number2/pxc3874633.pdf
4. V. V. Lakshmi, "Oil Spill Detection in Oceans using Threshold Segmentation and SVM classification," p. 4.
5. Z. Shao, S. Yang, F. Gao, K. Zhou, and P. Lin, "A new electricity price prediction strategy using mutual information-based SVM-RFE classification," *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 330–341, 2017.
6. M. Sahlgren, "An introduction to random indexing," 2005.
7. S. De Cnudde, D. Martens, T. Evgeniou, and F. Provost, "A benchmarking study of classification techniques for behavioral data," 2017.
8. C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Systems with Applications*, vol. 82, pp. 128–150, 2017.