Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-1 (2021) 37–42

On the relevance of clustering strategies for collaborative prognostics

Matteo Balbi*, Laura Cattaneo *, Domenico Daniele Nucera*, Marco Macchi*

** Department of Management, Economics and Industrial Engineering, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano, Italy - (e-mail:laura1.cattaneo@polimi.it)*

Abstract: The innovative concept of Social Internet of Industrial Things is opening a promising perspective for collaborative prognostics in order to improve maintenance and operational policies. Given this context, the present work studies the exploitation of historical and collaborative information for on-line prognostic assessment. In particular, while aiming at a cost-effective prognostic algorithm, with an efficient use of the available data and a proper prediction accuracy, the work remarks the relevance of an optimized clustering strategy for the selection of the useful information.

Keywords: Collaborative prognostics, data-driven prognostics, clustering, RUL prediction.

1. INTRODUCTION

A relevant challenge faced by the industry in this period is the minimization of machines' downtime and the costs related to it. This problem is strictly connected to the machines' failures and inefficiencies. The need to limit these events has promoted innovative capabilities of prognostics and health management, particularly aimed at estimating the time before the failure of a machine/equipment happens. The activity to accomplish this task is prognostics, which is aimed at estimating the so called Remaining Useful Life (RUL), that is the amount of time the machine/equipment will continue to perform its function according to design specifications (Zio, 2012). Prognostic methods are usually divided in two main classes: model-based and data-driven methods (Xia *et al.*, 2018). Due to the growing availability of sensors and the consequentially significative amount of data provided by the machines, data-driven methods are a hot topic in the research agenda (Cui, Kara and Chan, 2020). Considering the innovative concept of Social Internet of Industrial Things (SIoIT), whose underlying idea is the definition of a network of 'things' working socially (i.e. like a society) and cooperating to reach a common objective autonomously (Li and Parlikad, 2016), the promising vision of collaborative prognostics is potentially usable (Palau *et al.*, 2019) to exploit the available data from different assets: thus, machines (or, generally speaking, assets) can collaborate with each other to improve their maintenance and operational policies. The simplest way is to share data between different assets in the same network, thus offering the chance to provide the prognostic algorithms with a wider set of data for the RUL prediction. The improvement of the prognostic performances attained in these aforementioned studies confirmed the potential of sharing data among different assets, finally opening the possibility to exploit huge historical libraries of degradation paths up to the failure occurrence (in the reminder, respectively indicated as failures libraries and failures paths). The availability of such failures libraries shared among assets, opens new queries for what concerns how each failure path (in the library) can be used for a successful prognostic assessment. On one hand, it is possible to use all the available data within the prognostic algorithm, weighting the contribution of each path for the similarity with the online degrading path (Baraldi *et al.*, 2017; Cannarile, Baraldi and Zio, 2019). On the other hand, clusters of similar datasets can be defined, thus using just a limited set of information for the subsequent RUL prediction (Wang *et al.*, 2008; Loukopoulos *et al.*, 2019). Two criteria potentially in trade-off are relevant to make the choice: the higher prediction accuracy, reachable based on a more complete dataset, and the lower computational cost and time, reachable after filtering the available data before RUL prediction. As a matter of fact, the need of time saving when running a prognostic algorithm can be justified considering two main requirements: firstly, reducing the system response time is a crucial point for online RUL prediction; secondly, the expanding use of cloud computing systems calls for 'lean' algorithms, with limited computational resource consumption, as cloud computing services bill resources on a pay-per-use basis (Lacheheub, Hameurlain and Maamri, 2020).

Considering the introduced context, this work aims at defining a methodology to extract the most useful pieces of information from a library of failure paths for RUL prediction, leveraging on the concept of data clustering for collaborative prognostics. Concisely, it can be asserted that the experience gained by a fleet of assets during their lifecycles could be used to attain a reliable analysis regarding the RUL of an asset running on a degrading path, without wasting computational time in the evaluation of all the available datasets, including also the ones characterized by a low prognostic value. This time (and cost) saving approach can be compared to the aforementioned 'all data approach', in which all the failure paths are employed for RUL prediction. As far as the authors know, any work discussing and comparing these two different approaches cannot be found in the literature.

To accomplish this purpose, the prognostic algorithm applied is the Similarity Based Method (SBM), a data-driven model

10.1016/j.ifacol.2021.08.004

presented by (Wang *et al.*, 2008). The SBM assigns a similarity score to each historical failure path, thus predicting the RUL of the online monitored component as a weighted average of the historical RULs. This algorithm has been selected among others since it allows to evaluate the effect of each single historical path on the final RUL prediction. Building on this background, an innovative data-selection method equipped with a heuristic optimization algorithm is also proposed, with some proofs on its capability for a lower computational time with the same prediction accuracy.

The work is organized as follows. Section 2 describes the research objectives and the experimental testbed. Section 3 starts evaluating two possible clustering strategies, applying the most convenient one to compare with an all data approach. Section 4 presents the innovative data-selection method, together with its results in the experimental testbed. Section 5 finally summarizes the outcome of the work and its potential extension in future researches.

2. RESEARCH DESIGN

2.1 Research Objectives

This paper aims to exploit the collaborative prognostics by defining the most convenient set of failure paths to be used by a prognostic algorithm during an online RUL prediction. To reach this main objective, three intermediate goals are addressed, deployed in correspondent research steps. At first, the clustering approach is evaluated through the comparison of two possible clustering strategies: the 'hard clustering' and the 'adaptive clustering'. On one hand, the hard clustering relies on predefined groups of similar failure paths, assigning the online degrading path iteratively to the closest cluster. The closest cluster is chosen based on its current Euclidean distance to the historical paths, that is one of the commonly used distance measures. On the other hand, the adaptive clustering identifies at each iteration the subset of histories having the highest similarity with the online path, thus creating different clusters at different iterations. Once the most convenient clustering strategy is selected, the second step is to find out the best approach between the all data and the clustering one, both in terms of prediction accuracy and computational time (and cost). After assessing the overall performance of the clustering strategy, the work proposes a 'modified adaptive clustering' strategy that, through a data-selection method equipped with a heuristic optimization algorithm, shows its capability to improve the performances both in terms of effectiveness and efficiency.

2.2 Research Methodology

An experimental setting is developed to test the different approaches. In particular, the testbed simulates a set of machines whose data are generated by means of a flexible algorithm that can be used to replicate different scenarios. The algorithm is used both for the definition of an historical failures library and for the definition of the online path, whose RUL is evaluated iteratively every time a new degradation sample becomes available. The entire procedure can be summarized considering the following experimental steps.

Historical data generation: a user defined number of assets' histories characterized by up to four possible failure modes (FMs) are simulated. The degradation process is represented by means of up to four different features characterizing each FM. For what concerns the generation strategy, each feature is simulated sorting random variables of pre-defined distributions (normal, uniform, exponential, gamma). Figure 1 shows the result of a random run of the generation algorithm.

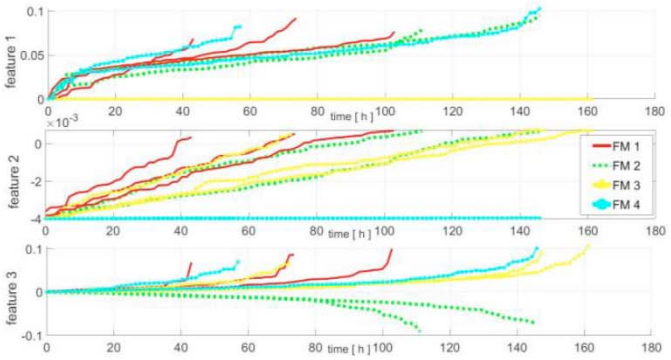

Figure 1: 10 histories for 3 dimensionless features describing 4 different FMs - one random run of the algorithm.

Definition and implementation of the diagnostic model: the slope of each feature is extracted and it is used to train a classification tree that will be employed to identify the incurring FM of the new online degradation dataset. This procedure corresponds to the diagnostic phase of a Condition Based Maintenance (CBM) program, whose function is to provide information that can be useful for building better models for prognostics (Jardine, Lin and Banjevic, 2006). Indeed, an effective evaluation of the incurring FM represents a crucial point to properly address the following steps.

Health Indicator (HI) construction: a HI is built to summarize different data from different sources (i.e. the four different features characterizing each failure path) (Lei *et al.*, 2018). In this work, a virtual HI is computed employing a multi linear regression as proposed in (Hu *et al.*, 2012). The method considers first of all the generated historical library and assumes a linear behaviour of the HI, ranging from 1 (start of the degradation) to 0 (end of the degradation). This allows to define, for each FM and for each feature, a weight (i.e. multiplier) to transform the raw generated features into the dimensionless virtual measure of degradation, i.e. the HI. Within this offline phase, the obtained array of weights is used to define the historical HI paths composing the historical reference library of failure paths. Figure 2 shows the HI paths of the 10 degradation paths up to failure occurrence, obtained considering the same random run of Figure 1.

Online data generation: A new set of features is simulated and is submitted time by time to the next two steps to mimic an online degradation process.

FM recognition and HI construction: the features' slope is extracted and processed by the diagnostic algorithm to recognise the incurring FM. Then, accordingly to the identified FM, the weights corresponding to the incurring FM are used to translate the raw features into the new online HI path.

Prognostic assessment: the online path is processed by the SBM, which assigns a similarity score to the paths composing the historical library for predicting the RUL.

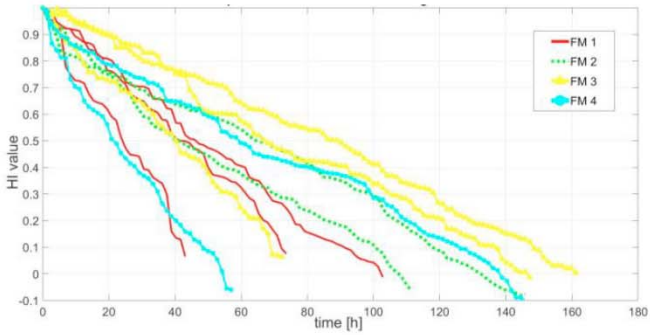

Figure 2: HI paths - generated historical library.

The described methodology and the subsequent clustering evaluations are implemented in MATLAB (R2020b).

3. COMPARISON OF CLUSTERING AND ALL DATA APPROACH

After having defined the experimental testbed, the work analyses which is the most convenient clustering approach, between hard and adaptive clustering. For the sake of clarity, the clustering strategy will be applied for selecting a proper subset of HI paths in the historical library that will be then used by the SBM in the prognostic phase.

The comparison between the clustering strategies is performed exploiting the online evaluation of the silhouette coefficient, which is an effective metric for assessing the clustering quality (introduced by Rousseeuw, 1987). In this work, the metric is used to evaluate if the online path has been well classified or not. Considering the prognostic perspective, the quality of the cluster assignment reflects the relevance of the histories included into the cluster itself, which will be then used for the online RUL prediction. In summary, the lower the clustering quality the higher is the prognostic potential of the excluded paths. This idea can be clarified considering the formulation of the silhouette coefficient related to the online path p:

$$sil(p) = \frac{v_p - u_p}{\max(v_p, u_p)}$$

In this expression, u_p is the mean distance of p from all the observations belonging to its cluster, while v_p is the minimum value among the distances w_{fp}, where w_{fp} represents the mean distance of p from all the observations of another cluster f.

For a direct comparison between the two strategies, the same random numbers are employed both for the definition of the historical library and of the new on-line path, whose silhouette coefficient is computed iteratively during the degradation process. Figure 3 shows the silhouette outcomes of the hard clustering corresponding to three generated paths. In this case the clusters computed on the historical HIs are defined at the beginning of the prognostic assessment and remain constant along the ongoing degradation process. The online HI will be assigned to one of them accordingly to the Euclidean distance at each time iteration.

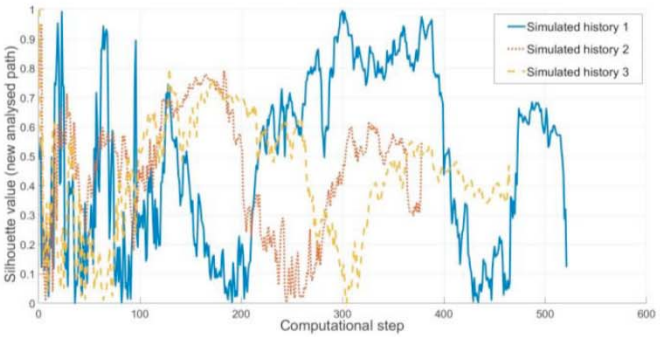

Figure 3: Online silhouette coefficient - hard clustering case.

A different result is obtained with the adaptive clustering strategy. In this work the adaptive clustering is implemented so as to arbitrarily consider 1/3 of the available paths; this is a rough decision, that will be discussed in the next section of the paper, where the proposed method will be refined. To carry out the experiment through the silhouette evaluation, the historical library is divided, at every time step, into three distinct clusters. The first, said reference, cluster is the one including the set of failure paths that will be employed in the subsequent prognostic phase (1/3 of the available paths). The second cluster is composed by the first two paths that are not included into the reference cluster. The third cluster is composed by the remaining paths. The decision of dividing the dataset in three different groups is driven by the choice of evaluating the proposed clustering method by means of the silhouette coefficient. Indeed, using only two clusters would have led to a definition of v_p affected by the presence of paths really far from the online one, causing a high *sil(p)*. Exploiting a three-cluster strategy is coherent with the description of the prognostic potential of the paths that, as a matter of fact, could bring the highest value to the estimation of the RUL. Figure 4 displays the online silhouette values obtained employing the adaptive clustering strategy.

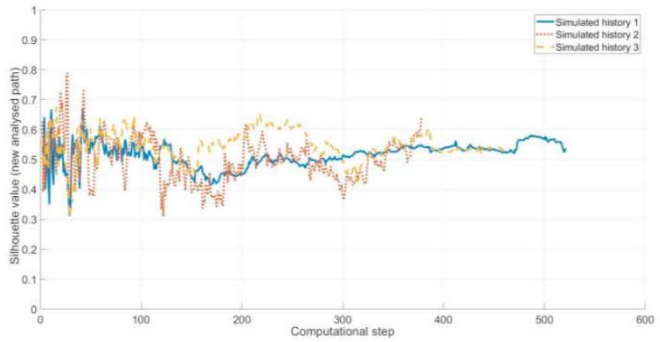

Figure 4: Online silhouette coefficient - adaptive clustering case.

Comparing the experimental results (Figure 3 and 4) is useful to underline the superiority of the adaptive clustering, as its online silhouette value is not characterized by the drops shown when the hard clustering is employed. The reduced silhouette values observed applying the hard clustering option are clear indicators that, in some specific computational steps, the considerable prognostic potential of some historical paths is excluded from the reference cluster and, therefore, will not be

used for the estimation of the RUL of the on-line path under analysis.

After having assessed the overall performance of the adaptive clustering strategy, this work aims at comparing the adaptive clustering with the all data approach, to define what is the most convenient method both in terms of prediction accuracy and in terms of computational time (and cost). For what concerns the experimental setting employed in this phase, four different FMs with similar failure paths and similar times to failure (TTF) are generated. This experimental setting defines a sort of limit case, in which the superiority of the clustering with respect to the all data approach or vice versa is not obvious. In fact, employing the all data approach in this context has been justified by different studies with the reasonable assumption that each dataset can bring value to the prediction (Baraldi *et al.*, 2017; Cannarile, Baraldi and Zio, 2019). On the contrary, in other situations where the different failure modes are characterized by different models/TTF distributions, it is intuitive to cluster the available data into smaller groups, to filter out the useless information given by the heterogeneous datasets.

The prognostic outcomes of the two different strategies, considering one random run, are shown in Figure 5.

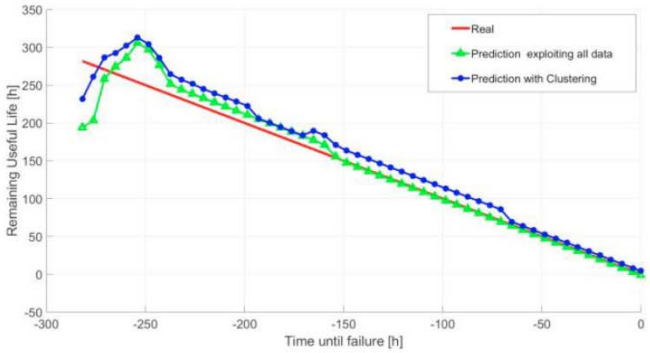

Figure 5: Prognostic results - comparing all data approach and adaptive clustering in the so called limit case.

The triangle-marked line represents the prognostic result obtained employing the all data approach while the dots-marked line corresponds to the adaptive clustering. After a first period where the results are different, the RUL prediction of the two different methods converges almost to the same value (overlaying on the real RUL known thanks to the simulated test-bed). The mathematical interpretation of this behaviour can be retrieved considering separately the two strategies.

In the case of the adaptive clustering, the initial estimate of the most similar paths could be biased due to an initial overlapping of the historical paths (Figure 2), leading to a wrong RUL prediction. Then, as a consequence of the ongoing degradation of the online path, the algorithm is able to identify which are the more convenient paths to be integrated into the reference cluster for the prediction. Therefore, step by step, the reference paths are evaluated and included into the prognostic algorithm, resulting in a reliable estimation of the RUL. On the other hand, considering the all data approach, the similarity score of the paths with a TTF comparable to the online path increases as long as the path develops, with a consequent decrease of the contribution of the other historical paths available in the library: the overall prediction accuracy increases gradually, obtaining a reliable estimate of the RUL when the path reaches the end of life.

As the prediction accuracy is similar for both the approaches, it is interesting to compare the two obtained computational times. The result of 9 independent runs of the algorithms processed on the same computer is proposed in Figure 6.

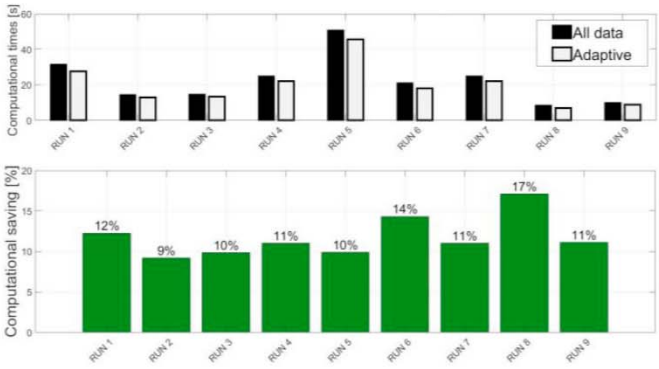

Figure 6: Computational times and computational saving results applying the adaptive clustering approach.

The figure shows how filtering the available information using an intermediate stage is beneficial for a faster computation of the RUL. In fact, applying the adaptive clustering allows to exclude from the prognostic algorithm the failure paths with a reduced prognostic value. Therefore, the preliminary computational effort required from the adaptive clustering approach to filter the available information allows an overall computational saving, thanks to a final lower number of histories processed by the prognostic algorithm. Given the comparable prediction accuracy and the computational savings due to the adaptive clustering approach, also in this limit experimental setting, the suggestion is to apply a pre-filtering stage on the historical library, with the purpose of including in the prognostic phase only the most meaningful pieces of information.

4. MODIFIED ADAPTIVE CLUSTERING

The good results obtained for the adaptive clustering strategy have encouraged the use of this method. Nevertheless, it is worth remarking that the adaptive clustering adopts a pre-filtering stage on the available historical library, programmed in order to exclude 2/3 of the available histories. Despite the good prognostic results obtained with this arbitrary choice, it could be expected that a deeper study of the selected dataset guarantees a performance improvement. The work therefore aims at optimizing the selection of the historical paths to be processed by the prognostic algorithm. The newly proposed method is driven by the requirement to solve this selection problem with an intuitive, effective and graphical method. This led to define the proposed algorithm as heuristic.

The work relies on the definition of a second filtering stage to investigate the effective prognostic value of the paths selected by the adaptive clustering. This proposed second filtering stage has been named "*silhouette filter*". In summary, the designed selection procedure is composed by two main steps: first, the Euclidean distances are used to arbitrarily select 1/3 of all the

available paths; then, the silhouette filter is applied to optimize on the first selection. The heuristic algorithm is summed up as in the reminder:

- Application of the first filter: given N_{tot} the number of historical paths within the library, evaluation of the Euclidean distance between the online path and the historical paths; therefore, selection of the n_{max} paths characterized by the lower values of distance, where $n_{max}=1/3\ N_{tot}$;
- Application of the silhouette filter: iteration of a number of paths n_{test} (between n_{min} and n_{max}) to compute the silhouette coefficient of the online path considering two principal clusters: the first cluster is composed by the n_{test} paths with the highest similarity with the online path; the second "dummy" cluster is composed by the path in position $n_{test}+1$ (i.e. it is the first excluded path from the first cluster);
- Selection of the configuration whose silhouette coefficient is the highest among the $n_{max}-n_{min}+1$ ones that have been tested; then, subsequent utilization of the obtained paths for the prognostic assessment;
- Iteration of the algorithm at each time step, to define the most convenient set of paths to forecast the online component's RUL.

Employing the silhouette value for this analysis allows to find out what is the most convenient number of failure paths to be used in the RUL prediction. The iterative procedure ranges from $n_{min}=3$ (considered as the minimum acceptable value to provide the algorithm with a sufficient level of experience) to $n_{max}=1/3\ N_{tot}$ (the value defining the number of paths selected by the first filtering stage). All in all, the maximum silhouette value defined during the iterations is an indicator of the best configuration, i.e. the configuration where the first excluded failure path has the lowest affinity with the cluster which embeds the online path.

To provide the reader with a meaningful example regarding the potential of this method, it is proposed an experimental setting where it is intuitive to understand the result of the silhouette filter. Figure 7 shows the available historical library and the new degradation path, which is marked with squares and represented at the end of its degradation process.

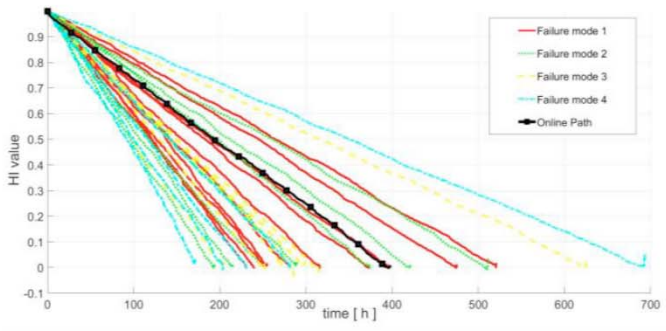

Figure 7: Heuristic optimization - proposed scenario described by dimensionless HI paths.

As it can be observed from the figure, there are four historical paths characterized by a considerable level of similarity with the online path. In this situation, it is intuitive to suggest to include these four histories within the prognostic algorithm, filtering out the information coming from the other historical paths. The goal of the proposed silhouette filter is to translate into a mathematical procedure this intuitive reasoning, thus offering the possibility to apply these principles also for more complex scenarios. Figure **8** describes the silhouette values corresponding to different clustering configurations obtained at different stages (thus, at different computational steps) of the proposed degradation process.

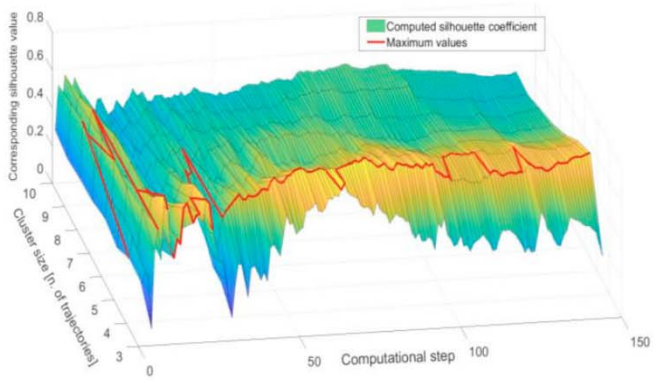

Figure 8: Heuristic optimization - number of paths selected by the proposed method.

The 3D representation of the designed selection process is useful to underline how the silhouette filter works. Indeed, after the initial steps in which, due to the overlapping of the available failure paths, the result is characterized by strong oscillations, the suggested cluster size reaches the expected value ($n_{test}=4$) in most of the analysed computational steps. In other words, instead of using $n_{max}=10$, the silhouette filter has optimized the cluster size. Considering a theoretical point of view, the paths excluded by the applied filter are the ones with a limited prognostic value and, therefore, their contribution on the subsequent RUL prediction is reduced. This consideration is supported by the prognostic results extracted from the proposed example which are described in Figure 9.

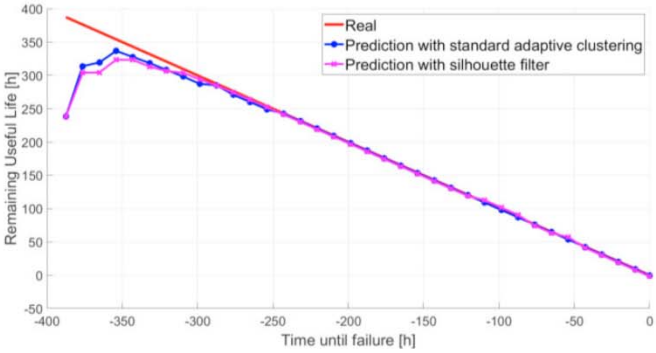

Figure 9: Heuristic optimization - prognostic results.

Figure 9 shows how the results of the adaptive clustering (line marked with dots) and of the modified adaptive clustering, with silhouette filter (line marked with crosses) are comparable. On one hand, the standard method filters 2/3 of the available information, relying on a limited library which

includes the four most significant paths. On the other hand, the modified method relies exclusively on the most significant paths, neglecting the contribution of the paths characterized by a reduced similarity score when the SBM is applied. Finally, Figure 10 shows the savings in terms of computational time obtained applying the proposed heuristic optimization. The results consider 9 independent runs of the algorithm, in which different experimental settings are applied. As it is shown in Figure 10, the silhouette filter adaptive clustering approach offers significant savings in terms of computational time.

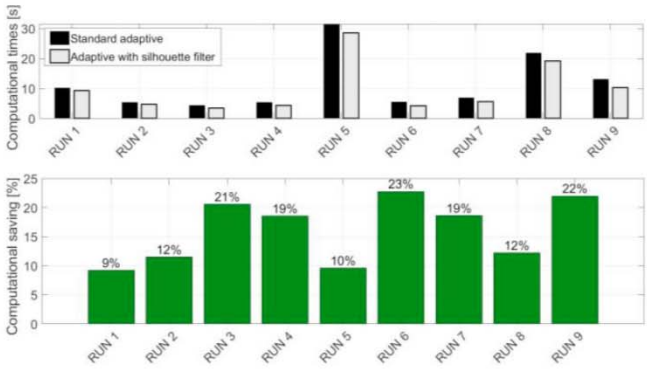

Figure 10: Computational times comparison and computational saving results applying the silhouette filter.

5. CONCLUSIONS

The availability of complete failures libraries in the context of collaborative prognostics calls for the investigation of which failures paths can be employed for a successful prognostic assessment. This work has demonstrated how to define the most convenient dataset to be employed in the RUL prediction. In particular, the good performances observed applying the suggested heuristic encourages to utilize the proposed double filtering method for an optimal selection of the historical data available from the complete failures libraries. The observed prognostic performances, and the considerable computational savings attained using the proposed clustering methods, justify the utilization of these approaches also in real industrial cases for two main reasons. First, the reduced time for a new prediction is an advantage in the prognostic field, since it reduces the system response time. Second, the lower requirement of computational resources reduces the expense for cloud computing services, in which the expense for the resources is determined on a pay-per-use base. This happens without losing the prediction accuracy. Therefore, the present work has obtained a promising outcome for a cost-effective deployment of collaborative prognostics. This could represent a starting point for the future definition of a working procedure suitable for a collaborative prognostic problem solving, built on a data-driven method in real industrial setting. Overall, the results can empower the development of the concept of SIoIT by introducing criteria for the cost-effective collaborative prognostics in fleets of assets dispersed in different working contexts and socially cooperating to reach the proper balance of prediction accuracy and cost/time trade-offs. These criteria have to be combined with the implementation of a proper information system to share maintenance data from different plants and assets, as reported in (Ardila *et al.*, 2020).

REFERENCES

Ardila, A. *et al.* (2020). XRepo-Towards an information system for prognostics and health management analysis. *Procedia Manufacturing*, 42, 146-153.

Baraldi, P. *et al.* (2013). Model-based and data-driven prognostics under different available information. *Probabilistic Engineering Mechanics*, 32, 66-79.

Cannarile, F., Baraldi, P. and Zio, E. (2019). An evidential similarity-based regression method for the prediction of equipment remaining useful life in presence of incomplete degradation trajectories, *Fuzzy Sets and Systems*, 36?, 36–50.

Cui, Y., Kara, S. and Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. *Robotics and Computer-Integrated Manufacturing*, 62,

Hu, C. *et al.* (2012). Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. *Reliability Engineering and System Safety*, 103, 120-135.

Jardine, A. K. S., Lin, D. and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance, *Mechanical Systems and Signal Processing*, 20, 1483–1510.

Lacheheub, M. N., Hameurlain, N. and Maamri, R. (2020). Resources consumption analysis of business process services in cloud computing using Petri Net, *Journal of King Saud University - Computer and Information Sciences*. 32, 408–418.

Lei, Y. *et al.* (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction, *Mechanical Systems and Signal Processing*. 104, 799–834.

Li, H. and Parlikad, A. K. (2016). Social Internet of Industrial Things for Industrial and Manufacturing Assets, *IFAC-PapersOnLine*. 49(28), 208–213.

Loukopoulos, P. *et al.* (2019). Abrupt fault remaining useful life estimation using measurements from a reciprocating compressor valve failure, *Mechanical Systems and Signal Processing*, 121, 359–372.

MATLAB, 2020. *version 9.9.0.1467703 (R2020b)*, Natick, Massachusetts: The MathWorks Inc.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, 20, 53-65.

Palau, S. A. *et al.* (2019). Collaborative prognostics in Social Asset Networks, *Future Generation Computer Systems*, 92, 987–995.

Wang, T. *et al.* (2008). A similarity-based prognostics approach for Remaining Useful Life estimation of engineered systems, *2008 International Conference on Prognostics and Health Management*, Denver, CO, pp. 1-6.

Xia, T. *et al.* (2018). Recent advances in prognostics and health management for advanced manufacturing paradigms, *Reliability Engineering and System Safety*, 178, 255–268.

Zio, E. (2012). Prognostics and Health Management of Industrial Equipment, *Diagnostics and Prognostics of Engineering Systems: Methods and Techniques*, 333–356. IGI Global, USA.