

Strategies to Face Imbalanced and Unlabelled Data in PHM Applications

Rafael Gouriveau, Emmanuel Ramasso, Nouredine Zerhouni

FEMTO-ST institute, UMR CNRS 6174 - UFC / ENSMM / UTBM, Automatic Control and Micro-Mechatronic Systems
Department, 24 rue Alain Savary, F-25000 Besançon, France
rafael.gouriveau@femto-st.fr

Accuracy and usefulness of learned data-driven PHM models are closely related to availability and representativeness of data. Notably, two particular problems can be pointed out. First, how to improve the performances of learning algorithms in presence of underrepresented data and severe class distribution skews? This is often the case in PHM applications where faulty data can be hard (even dangerous) to gather, and can be sparsely distributed accordingly to the solicitations and failure modes. Secondly, how to cope with unlabelled data? Indeed, in many PHM problems, health states and transitions between states are not well defined, which leads to imprecision and uncertainty challenges. According to all this, the purpose of this paper is to address the problem of "learning PHM models when data are imbalanced and/or unlabelled" by proposing two types of learning schemes to face it. Imbalanced and unlabelled data are first defined and illustrated, and a taxonomy of PHM problems is proposed. The aim of this classification is to rank the difficulty of developing PHM models with respect to representativeness of data. Following that, two strategies are proposed as pieces of solution to cope with imbalanced and unlabeled data. The first one aims at going through very fast and/or evolving algorithms. This kind of training scheme enables repeating the learning phase in order to manage state discovery (as new data are available), notably when data are imbalanced. The second strategy aims at dealing with incompleteness and uncertainty of labels by taking advantage of partially-supervised training approaches. This enables taking into account some *a priori* knowledge and managing noise on labels. Both strategies are proposed as to improve robustness and reliability of estimates.

1. Introduction

Data-driven Prognostics and Health Management (PHM) methods rely on the assumption that the statistical characteristics of data are relatively unchanged unless a malfunction occurs. These methods aim thereby at transforming raw monitoring data into relevant information and behavior models (including the degradation) of the system (Das et al., 2011), (Pecht and Jaai, 2010). Such methods are suitable for situations where it is hard to provide a mathematical model to replicate the behavior of physical system, or there is an absence of prior knowledge about the system. In other words, data-driven PHM methods can automatically learn to deduce complex and nonlinear relation among actual survival condition and measured condition monitoring information data, as they are trained to learn degradation from past examples (Dong, 2010). These methods are generally based on machine learning, artificial intelligence and pattern recognition tools.

The implementation phase of data driven PHM approaches has to go through important steps of learning and testing of the model (health assessment and/or prognostics model). Firstly, the model is tuned in order to learn behaviour of the system by pre-processed data (features) collected from degrading equipment, and secondly, the test phase uses learned model to predict the future condition (Figure 1). The modelling phase in itself can be split into two complementary steps: 1) pave the data space into areas of interest (data clustering), 2) build a behaviour model using the partition. According to this, the main limitation of data driven methods lies in the requirement of learning data: their performance is highly dependent on quality and quantity of data.

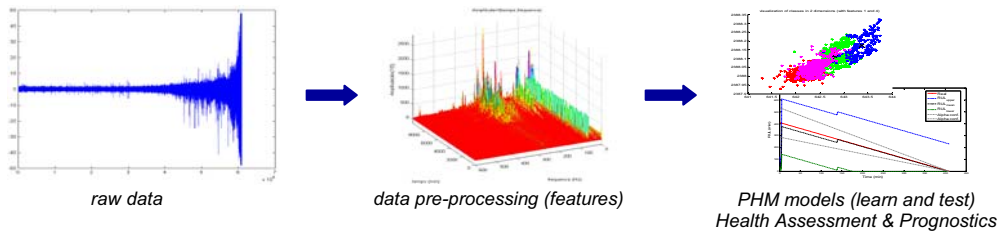


Figure 1: from raw data to PHM models

At least, two ill problems can be pointed out. First, learning dataset can be insufficient to accurately represent all possible states of the system (Figure 2 – left part). This is often the case in PHM applications where faulty data can be hard (even dangerous) to gather, and can be sparsely distributed accordingly to the solicitations and failure modes. Secondly, how to define health states and transitions between states if no prior knowledge is available, i.e. if data are unlabelled (Figure 2 – right part)?

According to all this, the purpose of this paper is to address the problem of "learning PHM models when data are imbalanced and/or unlabelled" by proposing two types of learning schemes to face it. Imbalanced and unlabelled data are first defined, and a taxonomy of PHM problems is proposed. The aim of this classification is to rank the difficulty of developing PHM models with respect to representativeness of data. Following that, two strategies are proposed as pieces of solution to cope with imbalanced and unlabeled data. The first one aims at going through very fast and/or evolving algorithms. This kind of training scheme enables repeating the learning phase as required in order to manage state discovery (as new data are available), notably when data are imbalanced. The second strategy aims at dealing with incompleteness and uncertainty of labels by taking advantage of partially-supervised training approaches. This enables taking into account some *a priori* knowledge and managing noise on labels. Both strategies are proposed as to improve robustness and reliability of estimates.

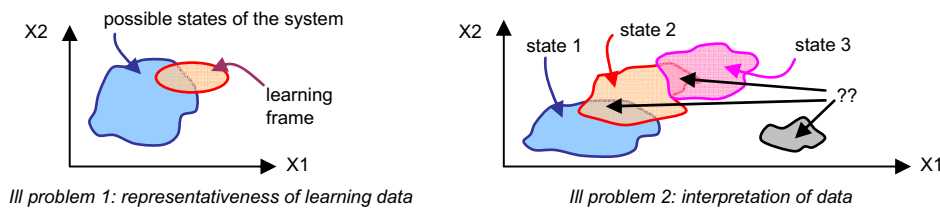


Figure 2: ill problems of learning data for PHM models

2. Taxonomy of PHM cases

2.1 Imbalanced and unlabelled data: terminology

Technically speaking, imbalanced data represents any dataset that exhibits an unequal distribution between its classes, i.e., cases in which the number of elements of one class severely outrepresents another. A taxonomy of imbalanced data was proposed in (He and Garcia, 1999) where authors distinguished between:

- "Intrinsic imbalances" that are directly related to the nature of the dataspace. This kind of imbalanced data can easily be found in PHM area, e.g. in nuclear power plant;
- "Extrinsic imbalances" that are related to time or storage. Even if the phenomenon is well balanced, data acquired can be imbalanced if a sensor has failed for example.

According to this, relative imbalance between classes or imbalance due to rare instances can result in a same problem for machine learning techniques and, whatever the type of imbalanced data, usual learning algorithms suffer from the disparity of samples in classes. Performances drop consequently (for instance, performances are also imbalance in between classes)...

Whilst training data represents generic knowledge that helps to capture inherent randomness of the data generating process, some specific knowledge can also be available to capture epistemic uncertainty due to lack of knowledge. In pattern recognition and machine learning algorithms, specific knowledge takes generally the form of prior information such as *labels*. In the context of PHM algorithms, labels can be viewed as a *ground truth* and can be encountered in two main cases:

- In the detection process: this process consists in discriminating between different possible functioning states, generally discrete. In this situation, a label associated to a training data point represents the real functioning state of that point (for example *healthy* or *broken*).
- In the prediction process: this process aims at estimating the remaining useful life of the system. When based on regression procedures, a label assigned to a training data represents the remaining useful life of this instance (for example *120 time units*).

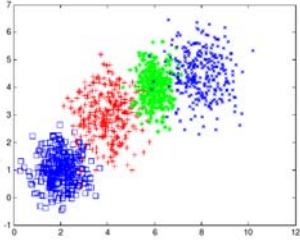
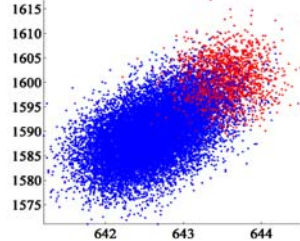
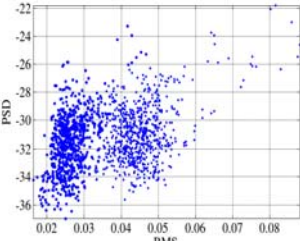
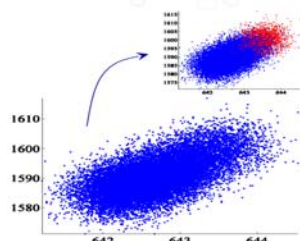
In both process, labels are used to improve estimates of the parameters used to discriminate between the states and predict the remaining useful life.

A dataset is called unlabelled when there is no prior/specific information. In that case, *unsupervised training* procedures are necessary to build PHM models. On the opposite, a labeled dataset is composed of a set of data points (time-series) plus prior information and one talks about *supervised training* data. In the sequel, we will expose two other cases: *semi-supervised training* and *partially-supervised training*.

2.2 Taxonomy of PHM cases accordingly to the data

According to the type of data, useful algorithms vary from an extreme to an other, and can require the user to *a priori* set some parameters to which performances are closely related (number of classes, distance measure, etc.). According to this, consider the following taxonomy of PHM cases (Table 1).

Table 1: Illustration of PHM cases – taxonomy accordingly to the available data

PHM Cases	Balanced Data	Imbalanced Data
Labeled Data	 <p>Case A - Modelling and estimates: easy</p>	 <p>Case B - Modelling and estimates: quite difficult</p>
Unlabeled Data	 <p>Case C - Modelling and estimates: quite easy</p>	 <p>Case D - Modelling and estimates: difficult</p>

▪ *Case A – PHM problem with balanced and labeled data.* Data of Case A (Table 1) are from a simulated example. In this case, data are well distributed (in quantity and in the space), and labels (states classes) are known. This can be the case for non costly components, i.e. components for which many failure experiments can be performed without compromising safety.

Such PHM issue can be easily addressed. Indeed, the current state of the component can be clearly identified and trajectories (behavior) are known. According to this, PHM algorithms can be built (they follow from traditional reliability modeling). However, is it a realistic case with respect to industrial constraints?

▪ *Case B – PHM problem with imbalanced and labeled data.* Data of Case B (Table 1) are extracted from the "Turbofan engine degradation simulation data set" (Saxena and Goebel, 2008). Blue points depict the normal mode (steady part) whereas red points are for faulty mode. One can note a strong overlap between classes with relative number of samples in each class. Also, labels are known since state classes are distinguished. This kind of data can depict sudden crack propagation phenomena, faults of an acquisition system (sensors), or excessive maintenance policies. In any case, these are explainable data sets.

In such PHM issues, since labels are known, local behavior models can be built. However, transitions between classes can be hard to catch: it requires the combination or adaptation of models. According to this, building PHM algorithms is quite difficult to perform (because of the dynamics of the behavior).

▪ *Case C – PHM problem with balanced and unlabeled data.* Data of Case C (Table 1) are extracted from experiments on PRONOSTIA platform (Nectoux et al. 2012). Data from two bearings with different loads are depicted. Data appears to be well distributed but labels are unknown. This kind of PHM problem can be encountered when large operating conditions or loads are considered, or if the quality of manufacturing is not constant.

The underlying structure of data is "depictable" and clustering of data can be done. This can lead to some problems of model parameterization and noise must be taken into account. Also, since no prior understanding of behavior is available, this has to be catch. However, efficient machine learning techniques for balanced data can be used, and developing a PHM algorithm for such cases is quite easy to perform.

▪ *Case D – PHM problem with imbalanced and unlabeled data.* Data of Case D (Table 1) are those ones from Case B without labels. One can guess a strong overlap between classes with relative number of samples in each class. Data are sparse, with rare instance and labels are unknown. This kind of data can depict cases where faulty states have never been met (new technologies, nuclear plants, etc.), or where the behavior is totally unknown like for multi-physics phenomena or multi-scales PHM problems.

In such PHM issues, the learning frame is poor which entails confidence problems. As for some examples, one should be able: to cluster data while distinguishing outliers from classes with few instances, to model transitions between classes thanks to combination or adaptation of models. However, since all situation have not been already met, all behaviors can not be modeled, and building PHM algorithms is difficult to perform (poor representativeness of data, misunderstanding of phenomena)

2.3 Pointing out a challenge

According to all above, at least two major issues can be pointed out:

- How to improve the performances of learning algorithms in presence of underrepresented data and severe class distribution skews? This is often the case in PHM applications where faulty data can be hard (even dangerous) to gather, and can be sparsely distributed accordingly to the solicitations and failure modes.
- How to cope with unlabelled data? Indeed, in many PHM problems, health states and transitions between states are not well defined, which leads to imprecision and uncertainty challenges.

The aim of next section is to discuss those challenges and to propose some strategies to cope with them.

3. Learning schemes to face imbalanced and unlabelled data in PHM applications

3.1 Evolving and/or fast algorithms: how to cope with balance of data

Consider Figure 3 to discuss the problem addressed. This can not be deeply presented in this paper but many PHM approaches based on artificial intelligent tools (neural networks NN, support vector machines SVM, hidden Markov models HMM, fuzzy inference systems FIS, etc.) have been proposed in literature. Nevertheless and even if it isn't always well pointed out by authors, all those methods are obviously dependent on the representativeness of the learning data. Indeed, in real case situations, data are not exhaustive and practitioners should be able:

- to distinguish outliers from transient;
- to manage "state discovery" (an extreme imbalanced case but practically useful);
- to cope with continuous data stream.

Also, learning PHM models can be time consuming and addressing those problems should be made in a time efficient manner. This is not the case when batch learning algorithms are required (like for some NN), or when optimization procedure are long (like in SVM approach). According to this, we propose two learning strategies as powerful candidates for PHM applications where imbalanced data are observed.

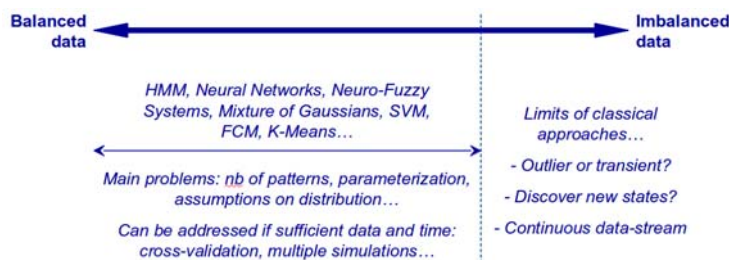


Figure 3: usefulness and limits of PHM approaches with respect to balance of data

- New health states will probably appear in the future (change in materials, drifts...). Thereby, one should consider upcoming data as potential behaviours to be learned. This can be achieved thanks to evolving systems with online algorithms. (El-Koujok et al. 2011) proposed an PHM models that starts from scratch and evolves (structure and parameters) as new data are gathered. This model provides practitioners with a tool that does not need the user to make assumptions on the structure or on initial condition for model building. The learning phase starts from scratch and the predictor evolves as data are gathered. Similarly, (Ramasso and Gouriveau, 2013) proposed a PHM model based on the combination of an evolving neuro-fuzzy predictor with an evidential classifier (based on belief functions). The approach appears to be very efficient since it enables to early estimate the failure instant, even with few learning data are available.
- An other way of dealing with state discovery is to imagine very fast learning schemes, i.e. algorithms that can be retuned as required as new data are available. (Javed et al. 2012) proposed a semi-complex extreme learning machine based on neural networks with complex activation functions to achieve health state monitoring and prediction. Experimental results show that with less complex network architecture, the proposed approach shows better accuracy performances, while reducing the processing time required (up to 130 times with respect with classical NN).

3.2 Partially supervised learning algorithms: how to cope with quantity and quality of labels

Consider Figure 4 to discuss the kind of problem that can follow from the quality and quantity of learning data. As state before, an underlying "data clustering" step is often required in PHM. It consists in gathering data points into similar regions of the feature space, also called clusters. This step is useful to decompose the estimation of the degradation model's parameters into simpler subproblems. For example, in Hidden Markov Models, the expectation-maximization algorithm allows to estimate the parameters of some probability densities which pave the feature space (Xing-Hui et al., 2010). Another example is the multi-modeling approach such as (Serir et al., 2012, 2013) which decomposes the feature space into regions for which one local model is used to estimate the evolution of the health indicators. Data clustering is also called "unsupervised" classification, meaning that the algorithm is able to estimate the clusters' parameters only based on the data obtained from sensors. In "supervised" classification methods, the data are accompanied by a ground truth that represents the real cluster. For real applications, knowing the ground truth means that the systems' health state is known for all data points (!). Another category of approaches are called "semi-supervised" approaches where only some data points are accompanied by the identity of the cluster. A problem can be pointed out: since supervised approaches seem to be irrelevant for real PHM approaches, how to deal with doubt on labels? According to this, we proposed to take advantage of what is called "partially-supervised" algorithms.

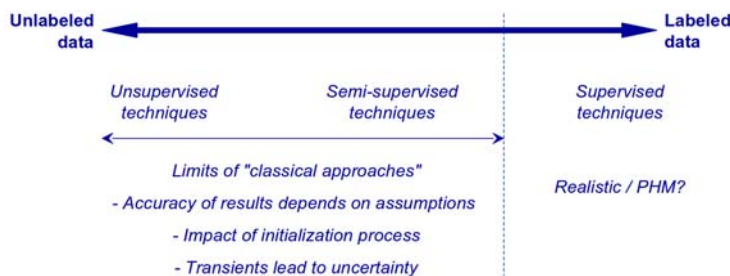


Figure 4: usefulness and limits of PHM approaches with respect to incompleteness & uncertainty of labels

- A recent approach called "partially-supervised" consists in considering that the ground truth can be known with uncertainty and imprecision (Côme et al. 2009) and covers semi-supervised, supervised and unsupervised training as particular cases. The formalism proposed in (Denoeux, 2013) is an extension of the previous work to consider uncertain and imprecise prior information in statistical model with latent variables. It was exploited for detection in (Ramasso et al. 2013a) and for prediction in (Ramasso et al. 2013b) using Hidden Markov Models. In all those works, the assessment of the partially-supervised learning schemes was performed with report to noise on labels. It was shown that the models are able to converge to relevant solutions even in the presence of noise on labels and with a small amount of data. Knowing precisely the true state in all data points is thus not necessary. It was also shown that the models provided accurate results when the labels were not "crisp", i.e. accompanied by uncertainty and imprecision. Besides, as demonstrated in (Denoeux, 2013), even though belief functions are used to encode the partial knowledge about the labels, the time/memory-consumption is not really influenced compared to usual models. It is all the more reduced than the labels are numerous and precise.

4. Conclusion

Data-driven PHM approaches are increasingly applied. However, accuracy and usefulness of learned PHM models are closely related to the availability and representativeness of data, as well as the interpretation of data. The aim of this paper is to address these problems by pointing out and discussing challenging topics for PHM modelling (within others learning challenges). Two aspects are considered: 1) how to deal with imbalanced data, i.e., with data whose relative number of instances in each class (each health state) evolves with time, and 2) how to deal with unlabeled data, i.e., data whose signification is not known by the user or at least poorly (with doubt). According to this, we propose two learning schemes to cope with incompleteness and imperfection of available learning data. The first one is based on evolving and fast algorithms that enable relearning PHM models as new data are available. The second one is based on "partially-supervised" learning algorithms that enable introducing doubt in PHM models.

Note that the problems considered in this paper are not the single ones to be addressed by PHM community. Indeed, others challenging topics like "robustness", "reliability", "verification" or "validation" of PHM models should be clearly stated by the research community as an area of required developments.

References

- Côme E., Oukhellou L., Denoeux T., Aknin P., 2009, Learning from partially supervised data using mixture models and belief functions, *Pattern Recognition*, 42-3, 334-348.
- Das S., Hall R., Herzog S., Harrison G., Bodkin M., 2011, Essential steps in prognostic health management, *IEEE Conference on Prognostics and Health Management*, Denver, CO, USA.
- Denoeux T., 2013, Maximum Likelihood Estimation from Uncertain Data in the Belief Function Framework, *IEEE Transactions on Knowledge Data Engineering*, 25-1, 119-130.
- Dong M., 2010, A Tutorial on Nonlinear Time-Series Data Mining in Engineering Asset Health and Reliability Prediction: Concepts, Models, and Algorithms. *Mathematical Problems in Engineering*, 22 p.
- El-Koujok M., Gouriveau R., Zerhouni N., 2011, Reducing arbitrary choices in model building for prognostics: an approach by applying parsimony on an evolving neuro-fuzzy system. *Journal of Microelectronics Reliability*, 51-2, 310-320.
- He H., Garcia E.A., 2009, Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21-9, 1263-1284.
- Javed K., Gouriveau R., Zerhouni N., Zemouri R., Xiang L., 2012, Robust, reliable and applicable tool wear monitoring and prognostic: an approach based on a Semi Complex Extreme Learning Machine (SC-ELM), *IEEE International Conference on Prognostics and Health Management*, Denver, CO, USA.
- Nectoux P., Gouriveau R., Medjaher K., Ramasso E., Morello B., Zerhouni N., Varnier C., 2012, PRONOSTIA: An Experimental Platform for Bearings Accelerated Life Test, *IEEE International Conference on Prognostics and Health Management*, Denver, CO, USA.
- Pecht M., Jaai R., 2010, A prognostics and health management roadmap for information and electronics-rich systems. *Microelectronics Reliability*, 50, 317-323.
- Ramasso E., Denoeux T., 2013a, Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions, *IEEE Trans. on Fuzzy Systems*, in revision.
- Ramasso E., Denoeux T., 2013b, Improved sequence prediction by Markovian generative models using partial knowledge encoded by belief functions, *In Int. Workshop on Partially Supervised Learning*.
- Ramasso E., Gouriveau R., 2013, RUL estimation by classification of predictions: an approach based on a neuro-fuzzy system and theory of belief functions, *IEEE Transactions on Reliability*, in revision.
- Saxena A., Goebel K., 2008, C-MAPSS Data Set, *NASA Ames Prognostics Data Repository*, <<http://ti.arc.nasa.gov/project/prognostic-data-repository>>, accessed, 28.01.2013
- Serir L., Ramasso E., Zerhouni N., 2012, Evidential Evolving Gustafson-Kessel Algorithm For Online Data Streams Partitioning Using Belief Function Theory, *Int. Jou. of Approximate Reasoning*, 53-5, 747-768.
- Serir L., Ramasso E., Nectoux P., Zerhouni N., 2013, E2GKpro: An evidential evolving multi-modeling approach for system behavior prediction with applications, *Mechanical Systems and Signal Processing*, in proof.
- Xing-Hui Z., Jian-She K., 2010, Hidden Markov models in bearing fault diagnosis and prognosis, *Second Int. Conf. on Computational Intelligence and Natural Computing (CINC)*, 2, 364-367.