INSTITUT FEMTO-ST

UMR CNRS 6174

Segmentation of CMAPSS health indicators into discrete states for sequence-based classification and prediction purposes

Version 1

Emmanuel Ramasso

Rapport de Recherche n° RR–FEMTO-ST–6839

DÉPARTEMENT AS2M and MEC'APPLI – January 13, 2016

Segmentation of CMAPSS health indicators into discrete states for sequence-based classification and prediction purposes

Version 1

Emmanuel Ramasso

Département AS2M and MEC'APPLI

PHM / T2DC

Rapport de Recherche no RR –FEMTO-ST–6839 January 13, 2016 (5 pages)

Abstract: CMAPSS (Commercial Modular Aero- Propulsion System Simulation) datasets have been initially developed for prognostics, forecasting and prediction, and available here: http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/. Those datasets represent run-to-failure time-series and have also widely been used for classification or clustering. For that, it requires a "ground truth" for performance evaluation. For homogeneity issue, a simple method is provided to automatically generate common "states/clusters/classes" which can be used by other researchers in benchmarking. Note that those states do *not* have real physical meaning concerning the behavior of turbofan engines. Matlab codes are also provided. The assumption is that (globally) monotonic health indicators have been extracted from the initial datasets.

Key-words: Time-series, sequence classification, prediction, prognostics, uncertain and noisy labels, health indicators and monitoring, CMAPSS datasets

FEMTO-ST Institute, AS2M and MEC'APPLI research departments
TEMIS, 24 rue Alain Savary, F-25000 BESANÇON FRANCE
Tel: (33 3) 81 40 28 01 – Fax: (33 3) 81 40 28 09 – e-mail: isabelle.gabet@ens2m.fr

Segmentation des indicateurs de santé des jeux de données CMAPSS en états pour la classification et la prédiction de séquences temporelles

Version 1

Résumé : Les jeux de données CMAPSS (Commercial Modular Aero- Propulsion System Simulation) ont été initialement proposés pour le pronostic et la prédiction. Ils sont disponibles à l'adresse suivante: http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/. Ces jeux de données ont par ailleurs été utilisés pour la classification et le partitionnement de séries temporelles, mais pour cela il est nécessaire de disposer d'une vérité terrain pour l'évaluation des performances. Pour des raisons d'homogénéité, une méthode relativement simple est présentée pour générer de manière automatique des états/classes/clusters pouvant être par la suite utilisés pour comparer des approches. Ces états n'ont *pas* forcément de lien avec la physique des systèmes de propulsion. Un lien vers les codes Matlab est fourni. L'hypothèse derrière l'utilisation de cette méthode est que les indicateurs de santé ont été extraits des données initiales et que ces indicateurs sont globalement monotones.

Mots-clés : Séries temporelles, classification de séquences, prédiction , pronostic, labels incertains et bruitéss, indicateurs et suivi de santé, CMAPSS

FEMTO-ST Institute, AS2M and MEC'APPLI research departments
TEMIS, 24 rue Alain Savary, F-25000 BESANÇON FRANCE
Tel: (33 3) 81 40 28 01 – Fax: (33 3) 81 40 28 09 – e-mail: isabelle.gabet@ens2m.fr

Segmentation of CMAPSS health indicators into discrete states for sequence-based classification and prediction purposes

Emmanuel Ramasso

January 13, 2016

Abstract

CMAPSS (Commercial Modular Aero- Propulsion System Simulation) datasets have been initially developed for prognostics, forecasting and prediction, and available here: http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/. Those datasets represent run-to-failure time-series and have also widely been used for classification or clustering. For that, it requires a "ground truth" for performance evaluation. For homogeneity issue, a simple method is provided to automatically generate common "states/clusters/classes" which can be used by other researchers in benchmarking. Note that those states do *not* have real physical meaning concerning the behavior of turbofan engines. Matlab codes are also provided. The assumption is that (globally) monotonic health indicators have been extracted from the initial datasets.

Key-words: Time-series, sequence classification, prediction, prognostics, uncertain and noisy labels, health indicators and monitoring, CMAPSS datasets

1 Description of the data

The turbofan datasets were generated using the CMAPSS (Commercial Modular Aero- Propulsion System Simulation) simulation environment that represents an engine model of the 90,000 lb thrust class [1, 2]. A number of editable input parameters was used to specify operational profile, closed-loop controllers, environmental conditions (various altitudes and temperatures). Some efficiency parameters were modified to simulate various degradations in different sections of the engine system. Selected fault injection parameters were varied to simulate continuous degradation trends. Data from various parts of the system were collected to record effects of degradations on 21 sensor measurements and provide time-series exhibiting degradation behaviors in multiple units.

These datasets possess unique characteristics that make them very useful and suitable for developing prognostic and health monitoring algorithms [3]: Multi-dimensional response from a complex non-linear system, high levels of noise, effects of faults and operational conditions, and plenty of units were simulated with high variability.

In the present document, the learning datasets of the four turbofan datasets are considered. The characteristics of the datasets are described in Table 1. It can be observed that dataset #1 is the simplest one with one operating condition (OC) and one fault mode. Datasets #2 and #4 are the most complex datasets with six OC and one or two fault modes. Dataset #3 presents two fault modes and one OC. The state-of-the-art results on those datasets until 2014 are presented in [3].

Datasets		#Fault Modes	#Conditions	#Train Units	#Test Units
Turbofan	#1	1	1	100	100
data from	#2	1	6	260	259
NASA	#3	2	1	100	100
repository	#4	2	6	249	248

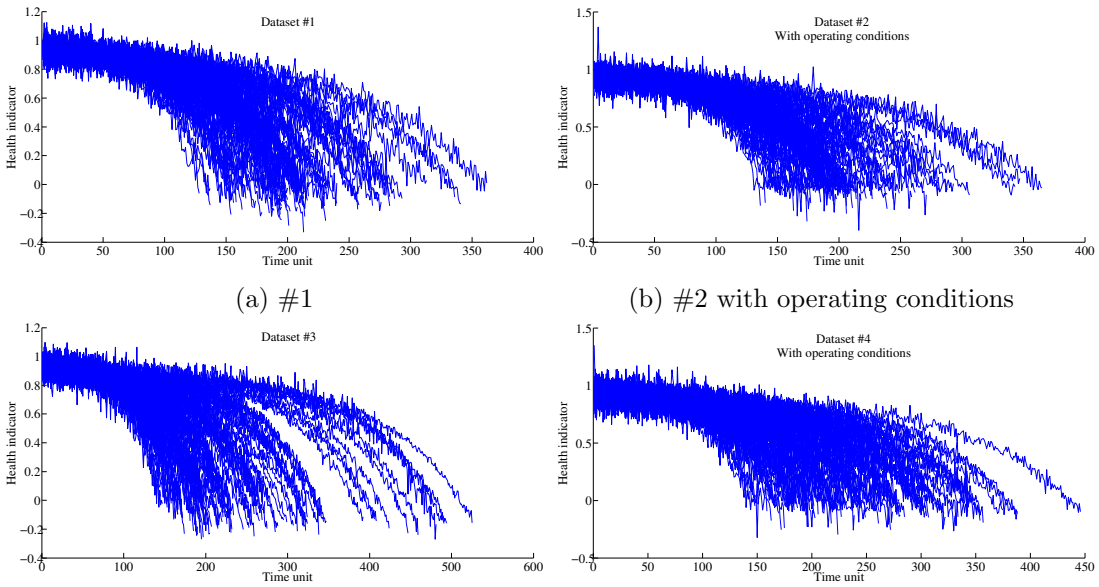
Table 1: Description of the turbofan degradation datasets available from NASA repository.

2 Getting health indicators

Each dataset is made of a certain number of trajectories with different length and 21 sensor measurements, with a total amount of about 700 training degradation trajectories (sum of values in Table 1). From sensor measurements, a health indicator is built for each trajectory as proposed in [4]. The method is based on Wang's approach [5]. A HI is a 1D signal which takes values in $[0, 1]$ (0 for failure state, 1 for healthy state) and if possible rather monotonic.

The health indicators (HI) for all trajectories and all datasets are depicted in Figure 1. We can observe that datasets #2 and #4 are made of about 500 trajectories with high variability in terms of noise and length as compared to dataset #1. These four sets of trajectories can be used for both training and testing data.

The codes to generate HIs are provided at:
http://fr.mathworks.com/matlabcentral/fileexchange/54866-rulclipper-algorithm-and-cmapss-health-indicators. The method is based on two strategies: Global which does takes operating conditions (OCs) into account, and local (considering OCs). OCs have an important impact on the quality of the HI in terms of noise level and degradation trend.



(a) #1

(b) #2 with operating conditions

(c) #3

(d) #4 with operating conditions

Figure 1: Evolution of the health indices for all engines in the four datasets.

3 Generation of the ground truth

Those datasets have been used for clustering or classification purposes in many papers as reported in [3]. We propose below a way to automatically generate three states from each trajectory in order to get a ground truth used to feed and compare methods of classification and prediction.

The process to get the segmentation into three states is depicted in Figure 2 and is available as a Matlab code at *http://fr.mathworks.com/matlabcentral/fileexchange/54808-segmentation-of-cmapss-trajectories-into-states*. The first 20% of a health indicator (HI) are used to estimate a linear regression model. State 1 (corresponding to the healthy state) starts at t=1 and ends when the absolute error between the estimation of the HI and the real HI is above 10%. State 3 (corresponding to the faulty state) starts when the absolute error falls below 10% (bottom right hand side figure) and ends at the last time unit. State 2 is finally found between both states. This process is repeated for all trajectories in all datasets and allows us to get one ground truth for each.

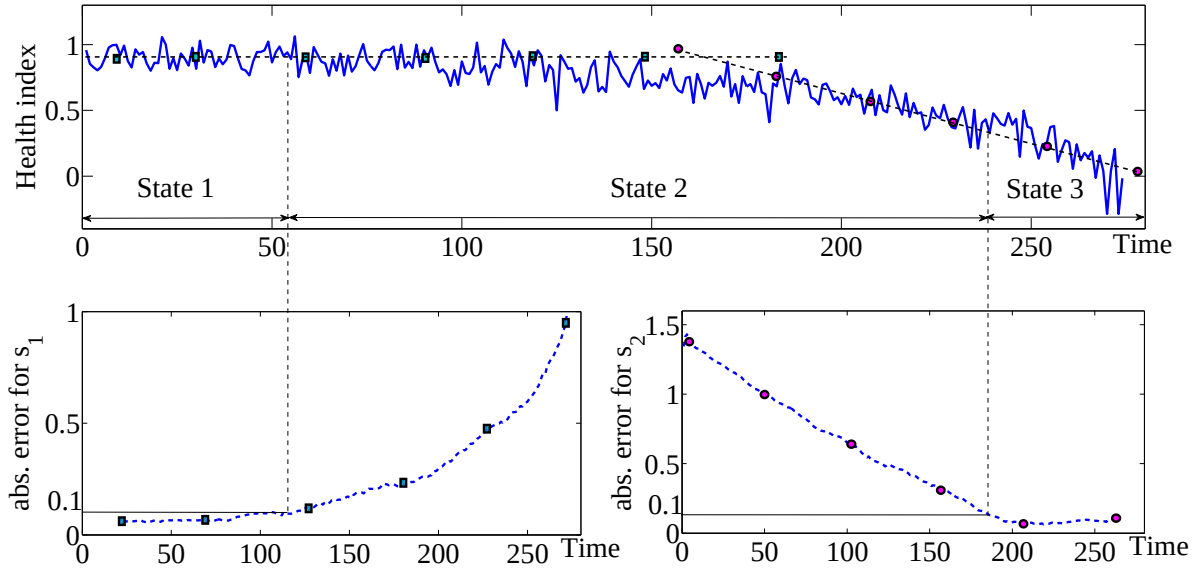


Figure 2: Automatic segmentation into three states.

The generation of the state sequences corresponds to a labeling which may be corrupted by errors due, for example, to the noise on the HI (impacted by OC and fault modes), and to the threshold (10% on the absolute error). In [6] we suggested an approach to incorporate noisy and uncertain labels (prior). Other approaches can be developed based on similar principle and using those data for comparison.

For models working with discrete variables, a quantization of the (continuous) HI can be performed to generate "symbols":

$$HI^{\text{discrete}}(t) \leftarrow \lfloor HI^{\text{continuous}}(t) \times N \rfloor \tag{1}$$

where N is a parameter playing a similar role as the number of clusters in usual quantization methods [7]. The influence of N should be studied according to the models.

Acknowledgement

This work has been carried out in the framework of the Laboratory of Excellence ACTION through the program "Investments for the future" managed by the National Agency for Research (references ANR-11-LABX-01-01). The authors are grateful to the Région Franche-Comté and "Bpifrance financement" supporting the SMART COMPOSITES Project in the framework of FRI2.

References

[1] D.K. Frederick, J.A. DeCastro, and J.S. Litt. User's guide for the commercial modular aero-propulsion system simulation (C-MAPSS). Technical report, National Aeronautics and Space Administration (NASA), Glenn Research Center, Cleveland, Ohio 44135, USA, 2007.

[2] A. Saxena, K. Goebel, D. Simon, and N. Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *International Conference on Prognostics and Health Management*, pages 1–9, Denver, CO, USA, 2008. IEEE.

[3] E. Ramasso and A. Saxena. Performance benchmarking and analysis of prognostic methods for CMAPSS datasets. *International Journal on Prognostics and Health Management*, 5(2):1–15, 2014.

[4] E. Ramasso. Investigating computational geometry for failure prognostics. *Int. Journal on Prognostics and Health Management*, 5(5):1–18, 2014.

[5] T. Wang. *Trajectory Similarity Based Prediction for Remaining Useful Life Estimation.* PhD thesis, University of Cincinnati, 2010.

[6] E. Ramasso and T. Denoeux. Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. *Fuzzy Systems, IEEE Transactions on*, 22(2):395–405, 2014.

[7] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

FEMTO-ST INSTITUTE, headquarters
15B Avenue des Montboucons - F-25030 Besançon Cedex France
Tel: (33 3) 63 08 24 00 – e-mail: contact@femto-st.fr

FEMTO-ST — AS2M / MEC'APPLI: TEMIS, 24 rue Alain Savary, F-25000 Besançon France
FEMTO-ST — DISC: UFR Sciences - Route de Gray - F-25030 Besançon cedex France
FEMTO-ST — ENERGIE: Parc Technologique, 2 Av. Jean Moulin, Rue des entrepreneurs, F-90000 Belfort France
FEMTO-ST — MEC'APPLI: 24, chemin de l'épitaphe - F-25000 Besançon France
FEMTO-ST — MN2S: 15B Avenue des Montboucons - F-25030 Besançon cedex France
FEMTO-ST — OPTIQUE: 15B Avenue des Montboucons - F-25030 Besançon cedex France
FEMTO-ST — TEMPS-FREQUENCE: 26, Chemin de l'Epitaphe - F-25030 Besançon cedex France

http://www.femto-st.fr