A Machine Learning Method Applied to the Evaluation of the Condition in a Fleet of Similar Vehicles

Pablo CalvoBascones

Institute for Research in Technology (IIT), Comillas Pontifical University – ICAI, Madrid, Spain.
E-mail: pcalvo@comillas.edu

Miguel A. Sanz-Bobi

Institute for Research in Technology (IIT), Comillas Pontifical University – ICAI, Madrid, Spain.
E-mail: masanz@comillas.edu

Chiara Brighenti

SATE. S.r.l. - Systems & Advanced Technologies Engineering, Venice, Italy.
E-mail: chiara.brighenti@sate-italy.com

Mattia Ricatto

SATE. S.r.l. - Systems & Advanced Technologies Engineering, Venice, Italy.
E-mail: mattia.ricatto@sate-italy.com

This paper presents a procedure for anomaly detection of temperatures in different key components of the power train in a fleet of similar vehicles. The anomaly detection is based on the characterization of the typical temperatures observed in the fleet of vehicles under all the working conditions that they develop. These typical temperatures are obtained by clustering methods and they are used as reference for identification of those vehicles where some abnormal behaviors, that can be symptoms of possible performance degradations or failures, are observed. The procedure uses data collected in real-time from the vehicle and they are used as inputs of a Self-Organized Map (SOM) able to discover the typical temperatures expected in their operation. The patterns obtained by the SOM cluster the vehicles according to similar behaviors concerning the temperatures observed at the different key points monitored. This offers a quick and effective view about the performance of each vehicle system respect to their reference temperatures obtained. Vehicles with untypical behaviors regarding the rest of vehicle fleet could suggest the existence of latent failures or degradations. Observing how each vehicle behavior shifts through the different neurons of the SOM, a prognosis can be made about the possible evolution of an anomaly detected. The paper includes some examples of application of the procedure used for the evaluation of the condition of the vehicle fleet.

Keywords: behavior patterns, normal behavior characterization, self-organizing maps, k-means, neural networks.

1. Introduction

Today industry is immersed in an important effort of digitalization (Binder et al. 2019, Vaidya et al. 2018). Also, the automotive industry is in this line since some years (Mastinu et al. 2019). More services can be offered to the driver and more information is available for monitoring the condition of the main components of a vehicle. This paper addresses this latter objective and analyses a particular case that was considered within a broader project performed by S.A.T.E. and IIT (see affiliations of the authors) for a renowned vehicles OEM (Original Equipment Manufacturer) for the development of tools for the condition monitoring of the key elements of an industrial vehicle in order to detect as soon as possible abnormal behaviors than can end in an undesired unavailability of the vehicle. Also, this type of strategy allows a better planning of maintenance and associated resources. In particular, this paper will describe a method to check the consistency of all the temperatures measured in a vehicle under several working conditions. If an abnormal value of a temperature is observed with respect to its expected value taking into consideration its consistency with the other temperatures, an incipient fault could be present that would have to be investigated. Machine learning techniques (Bonaccorso, 2017, Goodfellow et al. 2016) will be used to reach this goal.

This paper is organized as follows. Section 2 describes the objective of the study carried out. Section 3 presents the different variables of temperature monitored. Section 4 presents the method for building a model characterizing the patterns of temperatures observed. Section 5 describes the procedure used for anomaly detection and some examples. Finally, section 6 presents the main conclusions of the developed study.

Proceedings of the 30th European Safety and Reliability Conference and
the 15th Probabilistic Safety Assessment and Management Conference
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio
Copyright © ESREL2020-PSAM15 Organizers.*Published by* Research Publishing, Singapore.
ISBN: 978-981-14-8593-0; doi:10.3850/978-981-14-8593-0

2. Objective

The objective of this paper is the analysis of the coherence of temperatures measured in several key points of a powertrain of a vehicle. The idea is to discover as soon as possible abnormal values of temperatures that do not correspond to those expected for the observed operation conditions of the vehicle. The detection of an inconsistent temperature related to another observed is important in order to detect anomalies or degradations in components, but also for knowing the main focus of thermal problems that can have potential severe consequences on the operation conditions of other components of the vehicle. The detection of anomalies is also useful when different vehicles of a fleet are compared, so the vehicles that are working under different conditions than others can be detected, avoiding a degradation of conditions in the components affected and the function that they are carrying out.

The objective described is reached through the developments of the following consecutive tasks:

(i) Identification of temperatures to observe and to characterize.

(ii) Model building for pattern discovery of the temperatures observed.

(iii) Use of the patterns discovered for anomaly detection

They are described in the next sections of the paper.

3. Identification of temperatures to observe and to characterize

The study presented in this paper is based on several temperatures measured in real-time during the operation of each vehicle of the fleet. The list of measurements used is the following:

1. Intake manifold air temperature
2. Coolant temperature
3. Air intake temperature (upstream turbo compressor)
4. EGR cooler temperature
5. Air charger intercooler downstream temperature
6. Urea temperature
7. Exhaust gas temperature T1 (upstream DOC)
8. Exhaust gas temperature T2 (downstream DOC)
9. SCR upstream temperature
10. SCR downstream temperature
11. Environmental temperature

Where EGR, DOC and SCR refer respectively to the Exhaust Gas Recirculation, the Diesel Oxidation Catalyst and the Selective Catalytic Reduction process.

All these temperatures were recorded in °C at a sampling rate of 1 s.

4. Model building for pattern discovery of the temperatures observed

The mode to build has as objective the assessment of the temperature sensors consistency using the characterization of the combination of values typically observed in the temperatures during operation.

The main steps of the model building process are the following:

1. Selection of a data training set corresponding to time periods of normal operation of the vehicles. This will be used to obtain patterns of typical temperatures.
2. Model building. The objective of this step is to obtain a general model of temperature patterns for the whole fleet. This general model is based on the training data set previously selected.

The next sub-sections will describe these steps and the results obtained.

4.1 *Selection of a training set of data for model training*

The objective of this task is to obtain a dataset from which temperature patterns will be discovered. It is necessary to perform a pre-processing task of basic analysis of the data available in order to guarantee its reliability. In this way, for example, the chosen dataset must not include outliers, empty values for some or several variables, or continuously constant values in dynamic variables.

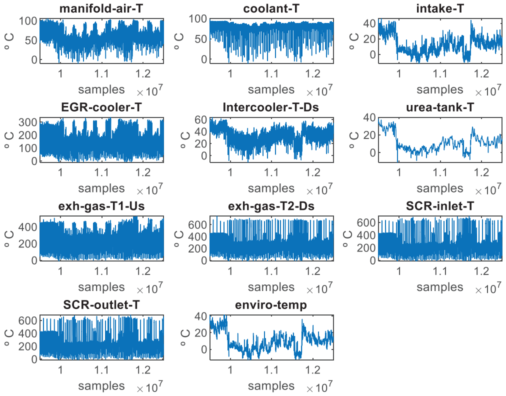


Fig. 1 - Training dataset of temperatures selected in a vehicle example.

After a previous pre-screening, the dataset selected included 3 million cases or samples and it was reduced to half of a million taking one sample out of 6 consecutive ones. At this point, it was verified that the relevant dynamics of the information was maintained after this process of data reduction. This accelerated the learning process maintaining the essence of the dataset from which the patterns will be discovered. As an example, Fig. 1 shows the evolution of the temperatures in a vehicle for a selected period of time representative enough of its normal behavior at several operation conditions. This is an example of a training dataset from which temperature patterns will be discovered.

4.2 *Model building*

The model developed is based on machine learning techniques (Bonacorso, 2017) for clustering. In particular, unsupervised clustering methods. The term “unsupervised” refers to the fact that no knowledge is required a priori for the classification and assignment of labels to distinguish the meaning of the profiles under the clusters. This is the case in this context because the idea is to find typical values of temperatures associated in several points of the vehicle under different working conditions without any a priori knowledge about what they have to be.

In this case, a Self-Organized Map – SOM (Kohonen, 2001, 2014) was used as an algorithm to discover different profiles of temperatures, but other algorithms could also be feasible. The idea behind the use of this algorithm is to classify the different samples of temperatures jointly observed in the typical operation of a vehicle, by a reduced number of patterns or clusters. This will facilitate to focus the attention of the huge amount of measurements collected into a small set of clusters.

The patterns to be obtained would jointly represent the most typical values of the 11 temperature variables selected for the temperature consistency analysis.

The number of patterns to use is difficult to decide a priori with accuracy. A well-known method used to suggest the optimum number of clusters to consider in a data set is the Elbow method (Kodinariya and Makwana, 2013). This method estimates the percentage of variance explained as a function of the number of clusters tested. The number of optimum clusters is found when testing the addition of a new cluster to the set of clusters does not add much new information.

In order to apply the Elbow method and to decide the most convenient number of clusters to use, another clustering algorithm named as k-means is applied (Reddy and Aggarwal, 2013). Fig. 2 presents the result of the application of the Elbow method and according to this figure, nine is the number of clusters suggested to use for an explanation of the variance of around 95%, which is a reasonable value. However, although nine is the number of clusters suggested, in this study twelve cluster were used to better guarantee the explanation of the variance of 95% or higher in the process of modelling.

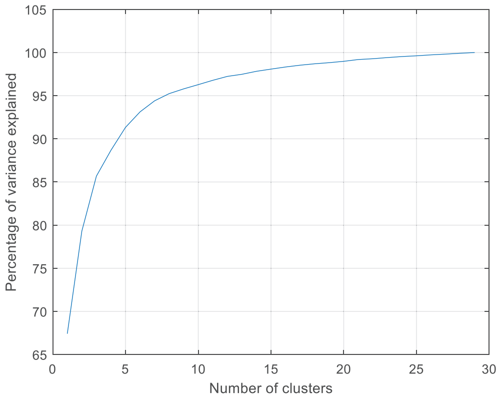
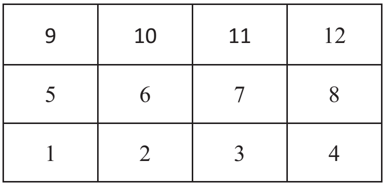


Fig. 2 - The Elbow method applied for the selection of the number of clusters for the model characterizing the normal behavior.

Once the training dataset was selected and the number of clusters to be used was decided, a model using the SOM (Self-Organized Map) algorithm was applied for discovering the 12 patterns or clusters tested. Even when k-means can suggest the patterns to consider, they were not used and in its place a SOM algorithm was developed for an easy interpretation of the clusters. The configuration of the SOM map was an architecture of 3x4 neurons with a hexagonal neighborhood region. MATLAB© was used as the tool for this study.



9	10	11	12
5	6	7	8
1	2	3	4

Fig. 3 - Topological representation of a SOM with dimensions 3x4. The numbers in the figure represent the numbering assigned to the neurons inside the map. This order is applied in the next three figures.

The characteristic of this map is the fact that neurons that are neighbor in the map will also have similar patterns. Referring to Fig. 3, which shows how the patterns are topologically placed in the map configured for the model, it can be said that patterns 1 and 2 are similar between them, and very different from patterns 11 and 12, which are at the opposite side of the map.

Fig. 4 shows the number of cases from the training dataset covered by each pattern / neuron. Since patterns are topologically placed as shown in Fig. 3, it is possible to observe from this Figure that the pattern 12 is the most frequent followed by the pattern 10. In general, the most part of the cases observed are in the row of neurons at the top of the map which are the most frequent patterns of behavior.

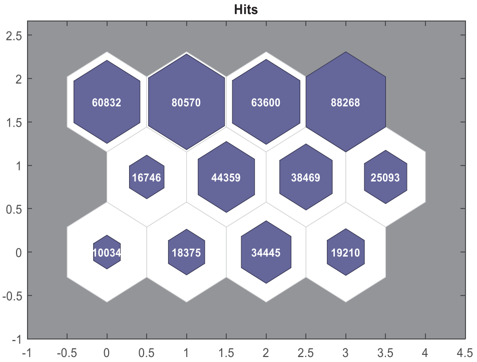


Fig. 4 - Hits or examples falling per each temperature pattern of a vehicle example.

A graphical representation of the patterns discovered is in Fig. 5. The numbers (1 to 11) in the x-axis shown for each graph correspond to the temperatures described in section 3 using the same order.

According to this figure the patterns discovered correspond mainly to different levels of the variables 4 (EGR cooler temperature) and 7 to 10 (Exhaust gas temperature T1 (upstream DOC), Exhaust gas temperature T2 (downstream DOC), SCR upstream temperature and SCR downstream temperature). These are the most relevant variables in this joint thermal analysis of temperatures in a vehicle. Also according to the machine learning technique used, which uses the auto-organization conception, it is possible to observe how the profiles of the patterns change smoothly along neighboring neurons from one corner of the map to the opposite in a diagonal line.

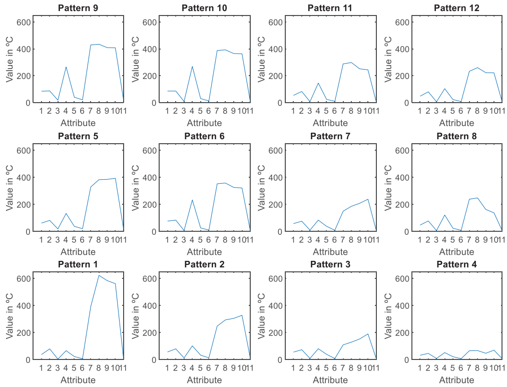


Fig. 5 - Patterns discovered in the temperature training set – Joint view of a vehicle example.

Fig. 6 is also useful to know better the patterns discovered. There, each subplot shows the contribution or sensibility of a specific attribute. The colors in each subplot represent from light to dark a higher contribution for this pattern to the model obtained. According to this, the Exhaust Gas Temperature T2 (Downstream DOC), the SCR Upstream Temperature and the SCR Downstream Temperature contribute in a similar way to explain the whole map. The same consideration can be made concerning the Environmental Temperature and the Air Intake Temperature, which have a very similar behavior explaining the map.

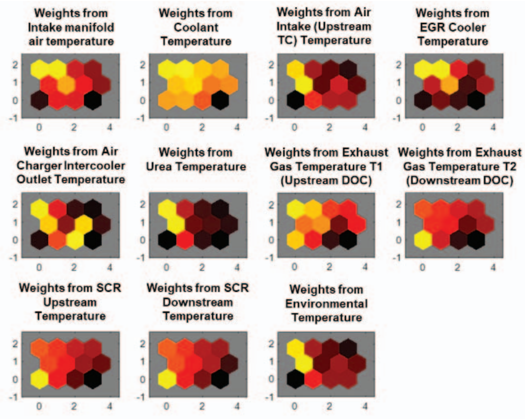


Fig. 6 - Influence of the input variables on the map obtained and patterns discovered in a sample vehicle.

After obtaining the model for the temperature patterns it is convenient to know the basic statistical characteristics of the cases that populate each pattern in order to know their distribution. Each pattern represented by the neurons in the SOM map has a centroid which represents the mean values of the attributes within the pattern and whose values were represented in Fig. 5. Also, Fig. 4 shows the number of cases from the training set belonging to each pattern. As an example, Table 1 presents the basic statistical values found for the cases of some patterns discovered and some input variables. This table is not complete due to the space required to present it.

Table 1 - Characteristics of some input temperatures used in the SOM map for some patterns in a vehicle example.

		Pattern 1	…	Pattern 11	Pattern 12
Attribute 1	mean	38,03	…	53,92	47,79
	std	12,76	…	8,64	7,86
	max	87,06	…	98,66	81,26
	min	12,36	…	13,46	17,36
Attribute 2	mean	79.56	…	82.58	80.33
	std	4.00	…	2.55	2.62
	max	92.96	…	91.46	91.46
	min	69.06	…	61.36	54.00
Attribute 3	mean	4.25	…	7.03	5.61
	std	9.02	…	7.66	8.25
	max	31.56	…	36.66	44.76
	min	-10.14	…	-11.74	-10.64
…	mean	…	…	…	…
	std	…	…	…	…
	max	…	…	…	…
	min	…	…	…	…
Attribute 11	mean	2.82	…	5.64	3.82
	std	9.40	…	7.90	8.44
	max	31.66	…	37.06	41.66
	min	-12.14	…	-12.64	-12.64

The process described previously corresponds to one case used for its illustration but in the same way it was applied to the rest of vehicles of the fleet obtaining similar results (see Fig. 7 where the patterns obtained from three vehicles are compared and it can be seen that they are quite similar). After this analysis it was decided to build a general model applicable to any vehicle of the fleet. A set of five training data sets from different vehicles were used selecting 500,000 samples from them. This was the training set used for obtaining the general model used as reference to discover anomalies. The procedure and results obtained for this general model were similar to those described before for a single vehicle. However the patterns obtained are more robust because they come from a diversity of vehicles preventing the loss of representativeness of the model.

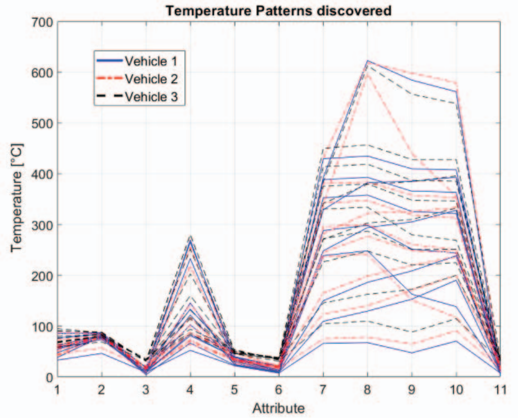


Fig. 7 - Comparison of patterns obtained from three different vehicles.

5. Use of the patterns discovered for anomaly detection

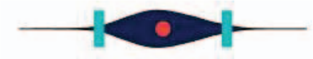
The anomaly detection method uses the general model that characterizes the normal behavior expected in the set of temperatures monitored in a vehicle. The method proposed is based on the three main elements that follow:

- Neuron attribute centroids: centroid values are the reference temperature patterns identified by the SOM algorithm.
- Standard deviation of each attribute within each neuron: for each attribute and each neuron, this is the standard deviation of the values taken by the attribute within that neuron in the training set.
- Probability density distributions of the temperature variables in the whole training set (global density functions) and within each neuron (local density functions).

Local probability distributions are obtained through a kernel density estimator, the bandwidth of which is obtained through the rule of thumb of Silverman (Silverman, B.W, 1986).Probability density distributions are discretized at equally-spaced values (100 samples for each local distribution) within the range [μ-4.5σ; μ+4.5σ]; where μ is the centroid value of the neuron for the temperature variable analyzed, and σ is its standard deviation. Discretizing density distributions with 100 equally-spaced samples is enough to compute the density values of new observations through the linear interpolation of a value and the equally-spaced samples. The range [μ-4.5σ; μ+4.5σ] ensures that the values at the probability distribution tails are 0; this range is justified due to the fact that a range equal to [μ-3σ; μ+3σ] includes 99.95% of the samples that make up the whole distribution according to the *3 sigma rule* of normal distributions (Friedrich Pukelsheim (1994)). This approximation assumes that the distributions of the temperature variables follow a normal distribution. This hypothesis is confirmed according to the analysis of these distributions.

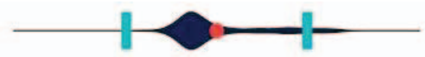
Three main types of probability density functions can be represented:

- **Normal distribution**: these density distributions belong to vehicles with similar variable values. Its graphic representation is:

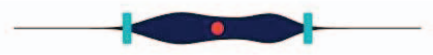


Where the red circle represents the neuron centroid μ, the two cyan rectangles represents the μ-2σ and μ+2σ values and the dark blue shape represent the distribution of the data. In this case the number of samples in correspondence of the centroids is the highest, as expected for a normal distribution.

- **Skewed distribution**: these density distributions belong to a set of vehicles in which most of them show equivalent variable values, but there are a lower number of vehicles with shifted variable ranges. Its graphic representation is:



- **Multi modal distribution**: these density distributions occur when there are two (or more) clusters of vehicles. Usually the centroid of neuron is located in between both distributions if the number of samples is the same in both clusters. Multi modal distributions usually appear when two or more different behaviors are under a same neuron. Its graphic representation is:



As an example, Fig. 8 shows the probability density functions of the cases represented by the twelve patterns of the general model obtained for the particular variable Exhaust gas temperature T1 (upstream DOC). It can be seen that the distributions are normal or slightly skewed.

So Fig. 8 shows the values of a single attribute of the SOM model and it can be seen how the structure of the SOM [3x4] determines how the neurons are distributed throughout the attribute axis.

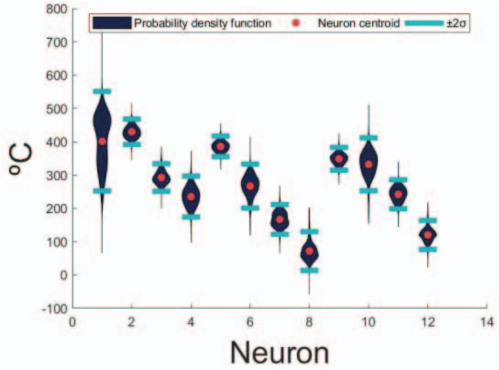


Fig. 8 - Probability density functions of the samples represented by the twelve patterns of the general model for the particular variable Exhaust gas temperature T1 (upstream DOC).

5.1 *Membership coefficient definition*

When a new observation of a temperature variable is collected, it has to be studied to detect any potential anomaly. This study requires to observe if the new observation is well represented by the corresponding probability density functions of the pattern to which it is assigned within the SOM. In order to assess the probability that the new observation belongs to such probability density functions, an index is defined in this paper, which is the "membership coefficient". This membership coefficient is the interpolated value of an observation within its corresponding discretized density function as shown in Fig. 9.

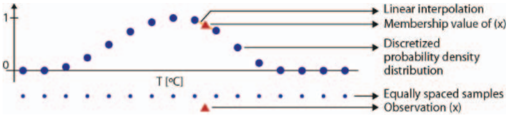


Fig. 9 – Example of how membership coefficient is computed for an observation (x) within a discretized probability density distribution.

Fig. 10 shows an example of how probability density functions (pdfs) are obtained for a generic temperature variable. In this example, a simple map with three neurons is considered. As can be seen, the global density function is computed over the entire training set, whereas the three local density functions are computed over three training subsets, made of the temperature samples assigned to the different neurons of the map. This figure shows also how the membership coefficient is obtained for a single observation, based on the normalized probability density functions.

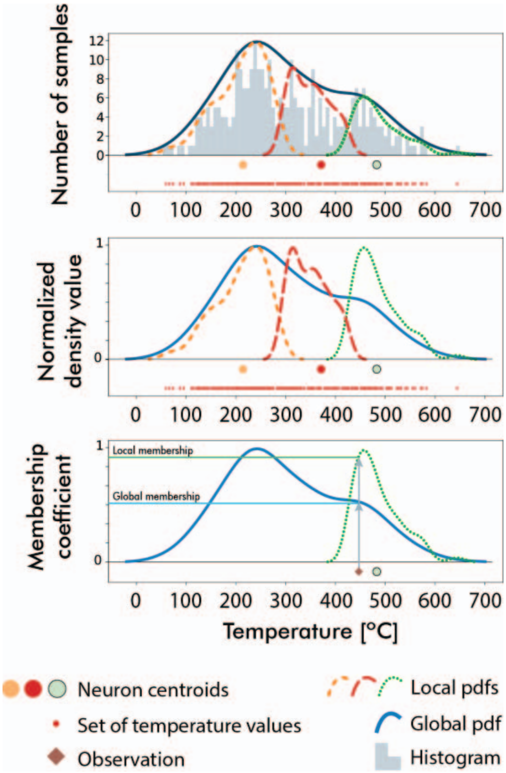


Fig. 10 – Example of probability density functions and membership coefficient computation for an observation.

Local and global membership coefficients are explained in the next section.

5.2 *Criteria defined for anomaly detection*

A set of criteria was set in order to evaluate how likely a new set of temperatures measured belongs to the reference model, which is the model built as described in the previous section. The criteria are based on three indicators, which are used to assess the temperature map:

- Local membership coefficient. This local indicator represents the location of an observation within the local probability density function of the closest neuron to which they belong. Very low membership values indicate that the new measured value is rarely present in the local distribution obtained from the training set.
- Distance from neuron centroid. This local indicator is an expression of the distance (quantified with regard to each variable standard deviation) between the centroid and the temperature value observed.
- Global membership coefficient. This global indicator represents the location of an observation within the global probability density function. Very low membership values indicate that the new measured value is rarely present in the global distribution obtained from the training.

Threshold values must be set for each indicator in order to differ normal conditions from warning and alert scenarios. When all the coefficients fall out of acceptance thresholds, an alert state is detected. In case that just some of them fall out of their acceptance thresholds, a warning scenario is detected.

5.3 *Examples of application of the anomaly detection*

In order to illustrate how this methodology was applied to the system previously described, the results obtained in the assessment of one of the variables of the model is shown in Fig. 11. This figure shows the evolution of a temperature signal during a certain period of time. In this period different scenarios took place: normal, warning and alert conditions. The threshold values set to define a normal condition were:

- Local and global membership coefficients have to be greater than 10^{-4}.
- Distance from the neuron centroid has to be lower than 3.

These conditions were adjusted according to the distributions obtained for the training set and based on some knowledge on the expected behavior of specific components made available by the experts. The bottom part of the figure shows how the system evolves from warning to alert conditions. It can be seen that the transition takes place fluctuating between both conditions. In those cases where the global membership coefficient is below its acceptance threshold, it means that the observation is not very frequent; in the case instead that the local membership coefficient is below its acceptance threshold, it means that the observation is not within the value ranges of the training set and, therefore, it is an anomalous value. Distance to centroid coefficient can be understood in a similar manner to the local membership coefficient. The results shown correspond only to a single value but they depend on the rest of signals of the model due to the fact that a neuron is active or not depending on the values of all the temperature variables.

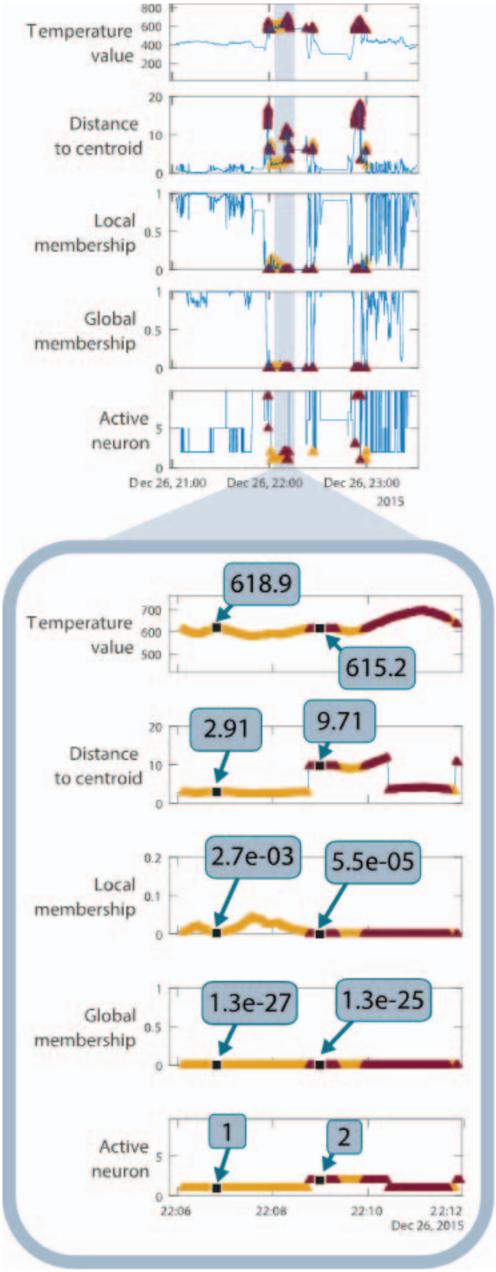


Fig. 11 Assessment of a temperature signal based on the temperature value, its distance to the centroid, its local and global membership coefficients and the neuron that is active for each observation. Warning conditions are marked in light color, alert conditions are marked in dark color

6. Conclusions

The methodology proposed for the assessment of behavior conditions based on indicators grounded on probability density functions and deviations from reference values determined through Self-Organized Maps is a novel strategy in the assessment of industrial components.

This methodology was applied in a fleet of multiple vehicles with the aim to improve the characterization of a normal behavior pattern used to assess future behaviors in the fleet of vehicles.

Warnings and anomalies detected were in most of the cases due to wrong sensors measurements or component malfunction, as confirmed by the association of the model results with the failures occurred to the vehicles and the maintenance interventions. This confirms that the methodology presented in this paper succeeds in the detection of normal, warning and alert conditions in a real scenario.

References

Binder, C., Neureiter, C. and Lastro G. (2019) Towards a model-driven architecture process for developing Industry 4.0 Applications. *International Journal of Modeling and Optimization 9-1*.

Bonaccorso, G. (2017) *Machine Learning Algorithms*. Packt Publishing

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. MIT Press

Kodinariya, T.M and Makwana, P.R. (2013) Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies 1-6*, 90-95.

Kohonen,T. (2001) *Self-Organizing Maps*. Springer

Kohonen,T. (2014). *MATLAB Implementations and Applications of the Self-Organizing Map*. Unigrafia Oy, Helsinki, Finland

Mastinu, G., Cadini, F., and Gobbi, M (2019) Industry 4.0 and Automotive 4.0: Challenges and Opportunities for Designing New Vehicle Components for Automated and/or Electric Vehicles. SAE Technical Paper 2019-01-0504.

Reddy C.K and Aggarwal, C.C. (2013) *Data Clustering*. Chapman and Hall/CRC

Vaidya, S., Ambad, P. and Bhosle, S. (2018) Industry 4.0 – A Glimpse. *Procedia Manufacturing 20*. 233-238.Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC. p. 45. ISBN 978-0-412-24620-3.

Friedrich Pukelsheim (1994) The Three Sigma Rule, The American Statistician, 48:2, 88-91, DOI: 10.1080/00031305.1994.10476030