HAL
open science

A framework for reconciliating data clusters from a fleet of nuclear power plants turbines for fault diagnosis

Sameer Al-Dahidi, Francesco Di Maio, Piero Baraldi, Enrico Zio, Redouane Seraoui

To cite this version:

Sameer Al-Dahidi, Francesco Di Maio, Piero Baraldi, Enrico Zio, Redouane Seraoui. A framework for reconciliating data clusters from a fleet of nuclear power plants turbines for fault diagnosis. Applied Soft Computing, 2018, 69, pp.213-231. 10.1016/j.asoc.2018.04.044 . hal-01988977

HAL Id: hal-01988977

https://hal.science/hal-01988977

Submitted on 8 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Applied Soft Computing 69 (2018) 213–231

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A framework for reconciliating data clusters from a fleet of nuclear power plants turbines for fault diagnosis

Sameer Al-Dahidi[a], Francesco Di Maio[a,*], Piero Baraldi[a], Enrico Zio[a,b], Redouane Seraoui[c]

[a] Energy Department, Politecnico di Milano, Milan, Italy
[b] Chair on Systems Science and the Energetic Challenge, Fondation EDF, Centrale Supélec, Paris, France
[c] EDF–R&DSTEP Simulation et Traitement de l'information pour l'exploitation des systèmes de production, Chatou, France

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 21 February 2018
Accepted 24 April 2018
Available online 28 April 2018

Keywords:
Fault diagnosis
Unsupervised ensemble clustering
Incremental learning
Cluster reconciliation
Fleet of nuclear power plants (*NPPs*)
turbines shut-down

ABSTRACT

When a fleet of similar Systems, Structures and Components (*SSCs*) is available, the use of all the available information collected on the different *SSCs* is expected to be beneficial for the diagnosis purpose. Although different *SSCs* experience different behaviours in different environmental and operational conditions, they maybe informative for the other (even if different) *SSCs*. In the present work, the objective is to build a fault diagnostic tool aimed at capitalizing the available data (vibration, environmental and operational conditions) and knowledge of a heterogeneous fleet of *P* Nuclear Power Plants (*NPPs*) turbines. To this aim, a framework for incrementally learning different clusterings independently obtained for the individual turbines is here proposed. The basic idea is to reconciliate the most similar clusters across the different plants. The data of shut-down transients acquired from the past operation of the *P NPPs* turbines are summarized into a final, reconciliated consensus clustering of the turbines behaviors under different environmental and operational conditions. Eventually, one can distinguish, among the groups, those of anomalous behavior and relate them to specific root causes. The proposed framework is applied on the shut-down transients of two different *NPPs*. Three alternative approaches for learning data are applied to the case study and their results are compared to those obtained by the proposed framework: results show that the proposed approach is superior to the other approaches with respect to the goodness of the final consensus clustering, computational demand, data requirements, and fault diagnosis effectiveness.
© 2018 Elsevier B.V. All rights reserved.

1. Introduction

In safety-relevant industries such as nuclear, oil and gas, automotive and chemical, fault diagnosis of Systems, Structures and Components (*SSCs*) is considered a critical task [1–3]. In particular, efficient fault diagnosis can aid to decide proper maintenance and, hence, increase production availability and system safety, while reducing overall corrective maintenance costs [4,5]. For these reasons, there is an increasing demand from industry for fault diagnosis techniques [6–9].

Generally, fault diagnosis techniques can be categorized into physics-based and data-driven [10,11]. Physics-based techniques use explicit physical models to describe the relationships between the causes that determine the *SSCs* behavior and the signal evolutions [11–13]. Several methods have been proposed and used for fault diagnosis in nuclear industry, such as observer-based methods, parity space methods, Kalman filters and parameter identification-based methods [14–16]. However, the complexity of the phenomena involved and the highly non-linear relationships between the causes and the signal evolutions may pose limitations on their practical deployment [11,13].

On the other hand, data-driven techniques are empirically built to fit measured process data [17–19]. For example, Artificial Neural Networks (*ANNs*), expert systems and fuzzy and neuro-fuzzy approaches have been successfully applied for fault diagnosis in the nuclear industry [20–22]. In this work, we focus on the development of a data-driven technique for fault diagnosis.

One attractive way forward for building effective diagnosis models is to consider the knowledge coming from the fleet of similar *SSCs* [3,23]. In the industrial context, the term fleet refers to a set of *P* systems that can share some technical features, environmental and operational conditions and usage characteristics. On this basis, three types of fleet can be envisaged: identical, homogenous and heterogeneous. Table 1 summarizes the types of fleet, their

* Corresponding author.
E-mail address: francesco.dimaio@polimi.it (F. Di Maio).

https://doi.org/10.1016/j.asoc.2018.04.044
1568-4946/© 2018 Elsevier B.V. All rights reserved.

Notation and list of acronyms

SSCs	Systems, structures and components
NPPs	Nuclear power plants
ANNs	Artificial neural networks
CSPA	Cluster-based similarity partitioning algorithm
RUL	Remaining useful life
FKNN	Fuzzy $\boldsymbol{K}$-nearest neighbours algorithm
ADASYN	ADAptive SYNthetic sampling approach
TOPSIS	Technique for order preference by similarity to an ideal solution
$\boldsymbol{H}$	Number of base clusterings
$\boldsymbol{j}$	Index of base clustering
$\boldsymbol{M}$	True number of clusters in the final consensus clustering
$\boldsymbol{C^j_{opt}}$	Optimum number of clusters of the $\boldsymbol{j}$-th base clustering
$\boldsymbol{P}$	Number of the *NPP* turbines of the fleet
$\boldsymbol{p}$	Index of the generic *NPP* turbine, $\boldsymbol{p} = 1, \ldots, \boldsymbol{P}$
$\boldsymbol{N_P}$	Number of shut-down transients of the $\boldsymbol{p}$-th *NPP* turbine, $\boldsymbol{p} = 1, \ldots, \boldsymbol{P}$
$\boldsymbol{i}$	Index of a transient, $\boldsymbol{i} = 1, \ldots, \boldsymbol{N_p}$
$\boldsymbol{Z}$	Number of signals of each i-th transient
$\boldsymbol{z}$	Index of the generic signal, $\boldsymbol{z} = 1, \ldots, \boldsymbol{Z}$
$\boldsymbol{T}$	Time horizon of the generic signal $\boldsymbol{z}$
$\boldsymbol{P^*_p}$	Optimum number of clusters in the final consensus clustering of the $\boldsymbol{p}$-th *NPP* turbine
$\boldsymbol{C_{min}}$	Minimum number of clusters in the final consensus clustering $\boldsymbol{P^*}$
$\boldsymbol{C_{max}}$	Maximum number of clusters in the final consensus clustering $\boldsymbol{P^*}$
$\boldsymbol{C_{Candidate}}$	Possible number of clusters in the final consensus clustering $\boldsymbol{P^*}$, $\boldsymbol{C_{Candidate}} \in [\boldsymbol{C_{min}}, \boldsymbol{C_{max}}]$
$\boldsymbol{P^*_{final}}$	The final reconciliated consensus clustering of the $\boldsymbol{P}$ *NPPs* turbines
DB	Davies- Boludin validity index
$\boldsymbol{N_{FF1/EE1}}$	Number of shut-down transients of FF1/EE1 *NPPs* turbines
$\boldsymbol{N_{aggregatedFF1,EE1}}$	Aggregated set of transients of FF1 and EE1 *NPPs* turbines
$\boldsymbol{P^*_{aggregatedFF1,EE1}}$	Optimum number of clusters in the final consensus clustering of the aggregated set of transients of FF1 and EE1 *NPPs* turbines
$\boldsymbol{m}$	Index of the generic consensus cluster of FF1, $\boldsymbol{m} = 1, \ldots, \boldsymbol{P^*_{FF1}}$
$\boldsymbol{\overline{\overline{Y}}^{e/f}}$	Vibrational measurements dataset of the $\boldsymbol{e/f}$-th transient of EE1/FF1
$\boldsymbol{e/f}$	Index of the generic shut-down transient of EE1/FF1, $\boldsymbol{e} = 1, \ldots, \boldsymbol{N_{EE1}}, \boldsymbol{f} = 1, \ldots, \boldsymbol{N_{FF1}}$
$\boldsymbol{\mu^m_{ef}}$	The similarity between $\boldsymbol{\overline{\overline{Y}}^e}$ and $\boldsymbol{\overline{\overline{Y}}^f}$ transients of the $\boldsymbol{m}$-th consensus cluster of FF1 *NPP* turbine
$\boldsymbol{\delta^m_{ef}}$	The pointwise difference between and transients of the $\boldsymbol{m}$-th consensus of FF1 *NPP* turbine
$\boldsymbol{y^{e/f}_{zt}}$	$\boldsymbol{t}$-th vibrational measurement of the $\boldsymbol{z}$-th vibrational signal of matrix $/\boldsymbol{\overline{\overline{Y}}^f}$
$\boldsymbol{C^*}$	Optimum number of clusters of the final consensus clustering and for the mean similarity values of each EE1 transient to FF1 consensus clusters
$\boldsymbol{\overline{\overline{X}}_{FF1}}$	FF1Training dataset matrix of FF1 *NPP* turbine
$\boldsymbol{\overline{\overline{X}}^*_{FF1}}$	Updated FF1 training dataset by ADASYN approach
$\boldsymbol{\overline{\overline{Y}}_{EE1}}$	Membership values of EE1 transients classification to FF1 consensus clusters
$\boldsymbol{\mu^m_e}$	Membership value of the allocating transient $\boldsymbol{e}$ of EE1 to the $\boldsymbol{m}$-th consensus cluster of
$\boldsymbol{K}$	Number of the nearest neighbors transients
$\boldsymbol{K_{min}}$	Minimum number of $\boldsymbol{K}$th nearest neighbors transients for the *FKNN* classifier
$\boldsymbol{K_{max}}$	Maximum number of $\boldsymbol{K}$th nearest neighbors transients for the *FKNN* classifier
$\boldsymbol{K_{Candidate}}$	Possible number of $\boldsymbol{K}$th nearest neighbors transients for the *FKNN* classifier, $\boldsymbol{K_{Candidate}} \in [\boldsymbol{K_{min}}, \boldsymbol{K_{max}}$
$\boldsymbol{K^*}$	Optimum number of $\boldsymbol{K}$th nearest neighbors transients used in *FKNN* classifier
CV	Cross validation analysis
$\boldsymbol{a^i}$	Average distance of the $\boldsymbol{i}$-th datum from the other data belonging to the same cluster
$\boldsymbol{b^i}$	Minimum average distance of the $\boldsymbol{i}$-th datum from the data belonging to a different cluster
$\boldsymbol{S^i}$	Silhouette value of the $\boldsymbol{i}$-th datum
$\boldsymbol{C_m}$	$\boldsymbol{m}$-th cluster in the final consensus clustering
$\boldsymbol{S_m}$	Mean Silhouette value for the $\boldsymbol{m}$-th cluster
$\boldsymbol{n_m}$	Total number of data in the $\boldsymbol{m}$-th cluster in the final consensus clustering
$\boldsymbol{SV_{C_{Candidate}}}$	Silhouette validity value at $\boldsymbol{C_{Candidate}}$, $\boldsymbol{C_{Candidate}} \in [\boldsymbol{C_{min}}, \boldsymbol{C_{max}}]$
$\boldsymbol{\overline{\overline{A}}}$	Adjacency binary similarity matrix
$\boldsymbol{\mu}$	Pairwise binary similarity value
$\boldsymbol{\overline{\overline{S}}}$	Co-association (Similarity) matrix
$\boldsymbol{S_{ij}}$	Pairwise similarity value between the $\boldsymbol{i}$-th and $\boldsymbol{j}$-th similarity values
$\boldsymbol{d_i}$	$\boldsymbol{i}$-th entry of the diagonal matrix $\boldsymbol{\overline{\overline{D}}}$
$\boldsymbol{\overline{\overline{D}}}$	Diagonal matrix with diagonal entries $\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots, \boldsymbol{d_N}$
$\boldsymbol{\overline{\overline{I}}}$	Identity matrix of size $\boldsymbol{N}$x$\boldsymbol{N}$
$\boldsymbol{\overline{\overline{L}}_{rs}}$	Normalized Laplacian matrix
λ	Eigenvalue of $\boldsymbol{\overline{\overline{L}}_{rs}}$
$\boldsymbol{S^m_e}$	Mean similarity value of transient e of EE1 to the whole transients of $\boldsymbol{m}$-th consensus cluster of FF1
$\boldsymbol{\overline{\overline{U}}}$	Eigenvectors of
$\bar{u}_{C_{candidate}}$	The $\boldsymbol{C_{Candidate}}$-th eigenvector of

characteristics and a selection of the most relevant research work performed in the past, making an effective use of fleet data:

In identical fleet, the systems might have identical technical features and usage, and work in the same environmental and operational conditions: knowledge derived from such fleet has been used for defining thresholds for anomaly detection [5], Remaining Useful Life (*RUL*) estimation [24] and technical solution capitalization [25,26] for any system identical to the fleet members.

In homogenous fleet, the systems might share some identical technical features that are influenced by similar environmental and operational conditions, but with few differences either on their features or on their usage: knowledge derived from this type of fleet has been used for developing diagnostics approaches for enhancing maintenance planning [27]. However, in a context where customized systems are common, these approaches may give poor results [3].

In heterogeneous fleet, the systems might have different and/or similar technical features, but with different usage under different

Table 1
Types of fleet, characteristics and most relevant literature.

Fleet type	*Characteristics*			*Objective of the work*
	Technical features	*Environmental and operational conditions*	*Usage*	
Identical	*Same*	*Same*	*Same*	*Anomaly detection* [5], *RUL estimation* [24] *and technical solution capitalization* [25,26]
Homogenous	*Same/Different*	*Same*	*Same/Different*	*Fault diagnostics and maintenance planning* [27]
Heterogeneous	*Same/Different*	*Different*	*Different*	*Fault diagnostics* [2,3,23].

environmental and operational conditions: this type of fleet can provide wider data and knowledge concerning the *SSCs* behaviour that are expected to reduce diagnosis uncertainty, and hence, improve the efficiency of the fault diagnosis task [2,3,23].

Most of the existing fleet-wide approaches for fault diagnosis treat only the information gathered from identical and/or homogenous fleets, rather than from heterogeneous ones [23]. In fact, the investigation on the benefit of utilizing the information of a heterogeneous fleet for fault diagnosis has been rarely addressed in the literature [23].

In this regard, the objective of the present work is to develop a framework for incrementally learning different turbine behaviours of a heterogeneous fleet of P Nuclear Power Plants (*NPPs*) turbines. The final goal is to summarize the data and knowledge acquired from the past experience of the fleet turbines operations into a final, reconciliated consensus clustering of the different turbines behaviors under different environmental and operational conditions (namely normal condition, degraded condition, abnormal condition and outliers).

In the context of fault diagnosis of an individual *NPP* turbine, the objective is to partition the N_p shut-down transients of the p-th plant, $p = 1, \ldots, P$, into M dissimilar groups (whose number is "a priori" unknown) such that transients belonging to the same group are more similar than those belonging to other groups. In particular, one can distinguish, among the groups, anomalous behaviors of the equipment and relate them to specific root causes [28–31].

The problem of grouping the operational transients of the turbine can be formulated as an unsupervised clustering problem aimed at partitioning the transient data into homogeneous "a priori" unknown clusters for which the true classes are unknown [30,32].

To this aim, an unsupervised clustering approach (sketched in Fig. 1) has been proposed by some of the authors for combining in an ensemble the clustering results of i) data representative of the turbine behavior, i.e., seven signals of the turbine shaft vibrations ($j = 1$ base clustering), and 2) data representative of the environmental and operational conditions that can influence the turbine behavior, i.e., nominal values of turbine shaft speed, vacuum and temperature signals ($j = 2$ base clustering) [32]. In brief, the approach is based on the combination of: 1) a Cluster-based Similarity Partitioning Algorithm (*CSPA*) to quantify the co-association matrix that describes the similarity among the two base clusterings (refer to Appendix A for more details); 2) Spectral Clustering embedding an unsupervised K-Means algorithm to find the final consensus clustering based on the available co-association matrix (refer to Appendix B for more details); 3) the Silhouette index to quantify the goodness of the obtained clusters by choosing the optimum number of clusters in the final consensus clustering as that with the maximum Silhouette value, i.e., such that clusters are well separated and compacted (refer to Appendix C for more details).

In this regard, the final ensemble clustering of the generic p-th *NPP* turbine comprises P_p^* clusters of shut-down transients, representative of different behaviors of the turbine that are influenced and explained by different environmental and operational conditions, among them some anomalous behaviors of the turbines can be identified [32]. The proposed approach has been applied to the shut-down transients of two different turbines of two different *NPPs* (coded as FF1 and EE1) of 149 and 116 multidimensional shut-down transients, respectively [32,33].

Due to the fact that the P plants of the fleet are highly standardized, some clusters representative of turbines operations and independently obtained for the individual plants might be similar (hereafter called the best matching clusters) and could be reconciliated into a unique cluster that would gather more information collected from multiple plants and, thus, is expected to be more reliable and robust.

More specifically, when a new dataset of N_{p+1} shut-down transients from the generic $p + 1$-th *NPP* turbine becomes available, the previously obtained ensemble clustering is updated based on the clusters identified independently for the transients of the $p + 1$-th *NPP* turbine.

The scope of this work is to propose a framework for identifying the best matching clusters among the plants: these will be reconciliated into a unique consensus cluster composed by the transients of the clusters independently obtained for the plants.

The proposed framework is validated on the two previously mentioned *NPP* turbines FF1 and EE1. The application of the framework leads to obtain a final, reconciliated consensus clustering P^*_{final} of 7 and 13 clusters representative of unique turbines operations of the FF1 and EE1 plants, respectively, and 3 consensus clusters representative of similar turbines operations of the plants (best matching clusters). The performance of the final reconciliated consensus clustering P^*_{final} is quantified in terms of clusters separation and compactness, by resorting to the Silhouette validity index ([34]; see Appendix C), *C-index* [35] and Davies-Boludin (*DB*) index [36]. The exploited knowledge of the turbines can, then, be retrieved for the purpose of, for instance, life tracking, health state estimation and fault diagnosis of a new *NPP* turbine.

For comparison, three other approaches are used to reconciliate the consensus clusters of the FF1 *NPP* turbine on the basis of the received information from the EE1 *NPP* turbine: 1) clustering of the aggregated shut-down transients of FF1 and EE1 *NPPs* turbines by the unsupervised ensemble clustering approach, 2) the inclusion of the EE1 transients into the FF1 ensemble clustering by resorting to Fuzzy similarity measure [37–39] and 3) the classification of EE1 transients by a supervised classifier, such as a Fuzzy K-Nearest Neighbours algorithm (*FKNN*) [40–42] trained on FF1 clustering. Results are discussed and compared with those obtained with the proposed approach: it is concluded that the proposed approach is able to update effectively the clusters of the FF1 *NPP* turbine on the basis of the received information from the EE1 *NPP* turbine, and that it is superior to the other approaches with respect to the goodness of the final consensus clustering, computational demand, data requirements, and fault diagnosis effectiveness.

Thus, the original contribution in this work is the development of a framework for incrementally learning the information brought by a heterogeneous fleet of different *NPPs* turbines based on the combination of:

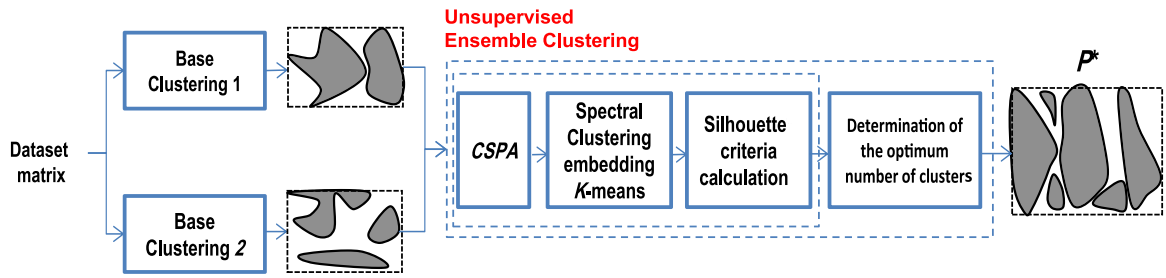

Fig. 1. The unsupervised ensemble clustering approach [32].

1) the unsupervised ensemble clustering approach [32], that overcomes the challenge to the existing clustering techniques by determining automatically the optimum number of clusters of the shut-down transients of each individual *NPP* turbine (which by most industrial applications, is not known "a priori"); the clusters that result are well separated and compacted (as measured by the Silhouette index [34]);
2) a reconciliation procedure for identifying the best matching clusters among the plants. The goodness of the final reconciliated clustering is quantified in terms of clusters separation and compactness.

It is worth mentioning that the dimensionality and required completeness of the datasets (that need signals representative of both environmental and operational conditions (i.e., turbine shaft speed, vacuum and temperature) and component behaviours (i.e., vibrations)) make, in this work, difficult to show the application of the framework to additional dataset from other industries, because of confidentiality constraints of such datasets.

The remaining of this paper is organized as follows. Section 2 illustrates the proposed framework for reconciliating the clusters of a fleet of industrial components for fault diagnosis. Section 3 and Section 4 describe how the proposed approach and three other alternative approaches are used for learning new data coming from a fleet of *NPP* turbines and updating the clustering results obtained by ensemble-clustering the transients coming from *NPPs*, respectively. Along with the description of the procedures, their application to the shut-down transients collected from a fleet of *NPPs* is shown. Finally, conclusions and perspectives are drawn in Section 5.

2. The framework for reconciliating the clusters of a fleet of industrial components

In this section, the framework for reconciliating the clusters of a heterogeneous fleet of P industrial components is proposed. The framework entails two steps and is sketched in Fig. 2:

Step 1: *Clustering the transients of a generic p-th component by the unsupervised ensemble clustering approach*. For the generic p-th component, the objective is to partition the N_p shut-down transients into dissimilar groups of transients representative of different component behaviors influenced by different environmental and operational conditions. To this aim, the unsupervised ensemble clustering approach of Fig. 1 (see Appendix A) has been set forth to build a consensus clustering P^* from the base clusterings:

1) $j = 1$: Clustering of data representative of the component behaviour (such as vibrations): the outcome of this is groups of transients representing different behaviours of the component, e.g., normal condition, degraded condition, abnormal condition and outliers,
2) $j = 2$: Clustering of data representative of the environmental and operational conditions that can influence the component behaviour (such as rotating speed, vacuum values, temperatures, pressures, etc.): the outcome of this is groups of transients representing different environmental and operationaconditions experienced by the component, e.g., a group might be characterized by high temperature values and low vacuum values.

The optimum number of clusters is selected among several candidates $C_{Candidate} = [C_{min}, C_{max}]$ based on the Silhouette validity index that measures the similarity of the data belonging to the same cluster and the dissimilarity to those in the other clusters (a large Silhouette value indicates that the obtained clusters are well separated and compacted ([34]; see Appendix C)).

Step 2: *Reconciliating the most similar consensus clusters obtained individually for each of the different plants*. To capitalize the added information of a new coming component (i.e., $p + 1$-th component) and, hence, to update the previous obtained consensus clustering $P_p^*$of the p-th component transients data, a reconciliation procedure is here proposed. The underlying approach is that of learning the novel information content of the new N_{p+1} transients without forgetting the previously acquired knowledge that is summarized in the P_p^*consensus clustering (as well shall see in Section 4). Firstly, the N_{p+1}transients have to be partitioned into groups representative of the $p + 1$-th component behavior under varying environmental and operational conditions of the new component as done in Step 1 for the p-th component. Once the consensus clusterings P_p^* and P_{p+1}^* of the two components are available, those composed by transients with similar behaviors are identified and reconciliated into unique clusters within the final ensemble clustering of the two plants $P_{p,p+1}^*$. The remaining clusters are left disjoint as they are representative of unique operational conditions of each component.

The incremental learning process and the enveloping reconciliation approach is repeated for all the components available in the fleet to get the final clustering P_{final}^*that resumes the characteristic behaviours of all the possible (available) components operating in as large as possible variety of environmental and operational conditions.

Once the final clustering P_{final}^* is obtained, the goodness of the final clusters identified is quantified in terms of their separation and compactness, as measured by internal validity indexes. These indexes evaluate the clustering results based on information intrinsic to the data itself, without resorting to any external information like true clustering results, which are not known "a priori" in most industrial applications [43]. In particular, we resort to the following three internal indexes:

- the Silhouette index ([34]; see Appendix C): it measures the similarity of the data belonging to the same cluster and the dis-

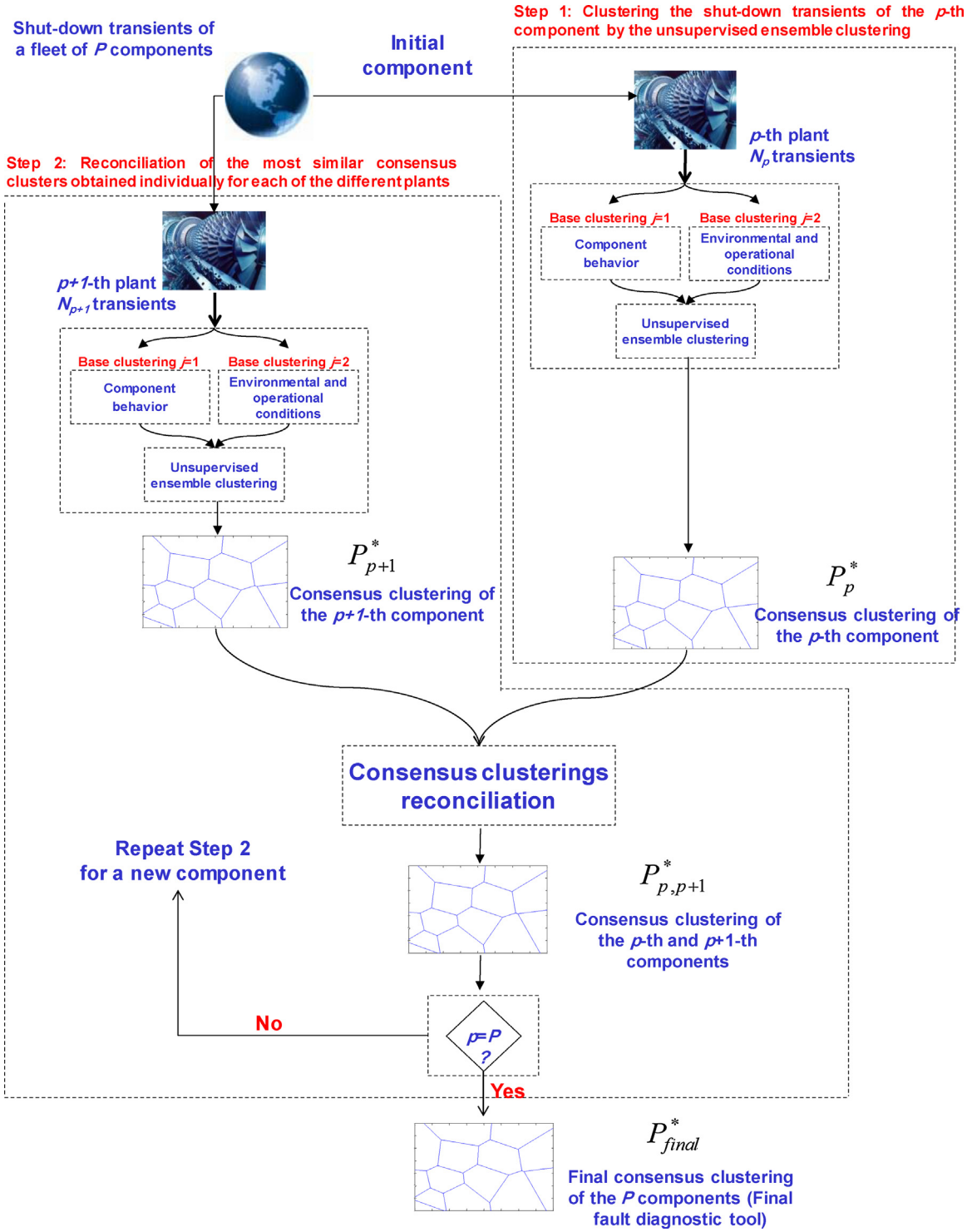

Fig. 2. The proposed framework for reconciliating the consensus clusterings for a fleet of P components.

similarity to those in the other clusters. The Silhouette index varies in the interval [−1,1] and should be maximized;

- the *C-index* [35]: it defines the ratio between the sum of within-cluster distances and the distances considering all the pairs of the instances. The *C-index* ranges in the interval [0,1] and should be minimized;
- the Davies-Boludin (*DB*) index [36]: it is based on the ratio of within-cluster and between-cluster distances. The *DB* index ranges in the interval [0,∞) and should be minimized.

Large Silhouette and small *C-index* and *DB* values indicate that the obtained clusters are well separated and compacted.

It is important to point out that there exist other clustering validity indexes, the so called external validity indexes, that evaluate the goodness of the obtained clusters with respect to a pre-specified structure (assumed to be known "a priori"), like false-positive, false-negative and classification error, etc. [43]. However, the calculations of these indexes are not feasible in this work due to the unavailability of the true clustering results.

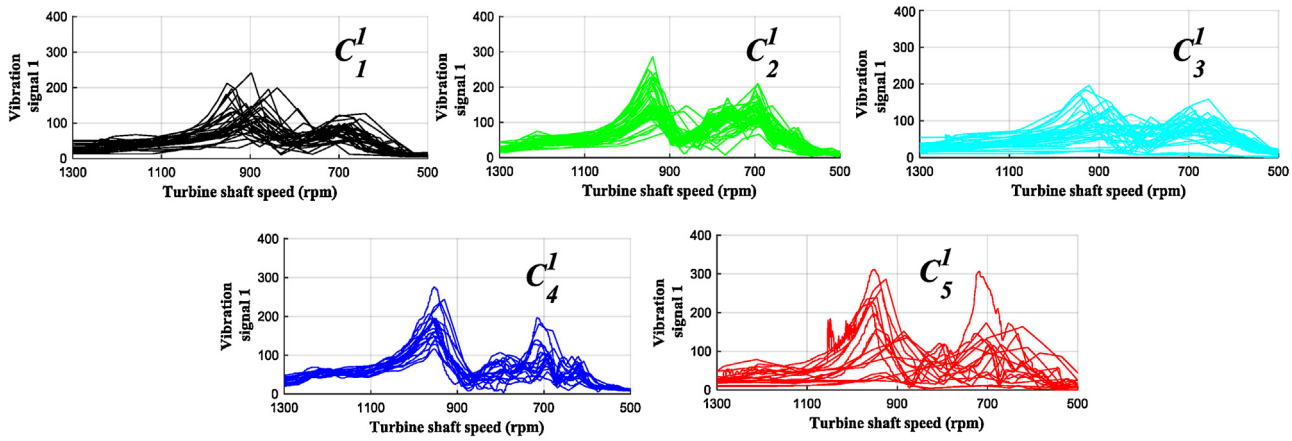

Fig. 3. Evolution of vibration signal 1 of the 5 obtained clusters of the $j = 1$ base clustering of FF1 *NPP* turbine and the corresponding turbine shaft speed values.

Table 2
Number of base clusters along with the optimum number of clusters in the final ensemble clustering of FF1 *NPP*.

$j = 1$ Base clustering.	$j = 2$ Base clustering	*Consensus clustering*
$C^1_{opt} = 5$	$C^2_{opt} = 4$	$P^*_{FF1} = 10$

In the following section, the proposed framework is applied to $P = 2$ *NPPs* turbines (FF1 and EE1).

3. Application of the reconciliation approach to two nuclear power plants turbines

In *NPPs*, the turbine is one of the most important and critical rotating machinery for generating a large efficiency and peak factor. Unexpected failures are, indeed, usually accompanied with large downtimes, high cost, as well as possible safety and environmental implications [11,44].

For these reasons, recognizing the health state of turbine (diagnostics) and predicting its future evolution (prognostics) are fundamentals to enable more reliable, economic and safer operation [45,46].

In this regard, we are looking for the failures of the turbine by investigating its shut-down transients since the turbine during shut-down transients are expected to provide better, more clear indications of the health state with respect to stationary conditions [30,47].

The application of the proposed framework presented in Section 2 is here presented with respect to a real industrial case concerning $P = 2$ *NPPs* turbines (i.e., FF1 and EE1) with different numbers of shut-down multidimensional transients of 149 and 116, respectively. Each i-th transient is a multidimensional transient in a $Z = 70$ dimensional signal space with a time horizon of $T = 4500$ time steps of 2.5 h each. It should be mentioned that throughout the paper, the values of the signals illustrated in the figures are given on an arbitrary scale due to confidentiality reasons.

3.1. Step 1: clustering shut-down transients of the FF1 NPP turbine by the unsupervised ensemble clustering approach

Starting from the available dataset of shut-down transients of the FF1 *NPP* turbine, different turbine behaviors explained by different environmental and operational conditions have been identified by resorting to the unsupervised ensemble clustering approach [32]. Table 2 reports the number of base clusters along with the optimum number of clusters obtained in the final consensus clustering P^*_{FF1}.

Table 3
Consensus results for $j = 1$ and $j = 2$ base clusterings.

$j = 1$ Base clustering	$j = 2$ Base clustering	*Consensus clustering*
C^1_1	C^2_3	P^*_1
	C^2_1	P^*_4
C^1_2	C^2_2	P^*_5
	C^2_3	P^*_6
	C^2_1	P^*_{10}
C^1_3	C^2_3	P^*_2
	C^2_{NaN}	P^*_3
	C^2_1	P^*_8
C^1_4	C^2_1	P^*_9
C^1_5	C^2_2	P^*_7

For the ease of clarity, Fig. 3 shows the evolution of vibration signal 1 of the $j = 1$ base clustering (5 clusters) and the corresponding turbine speed values. One can easily recognize that on one side the functional behaviors of transients belonging to clusters 1–4 (C^1_1, C^1_2, C^1_3, and C^1_4) are similar but with some peculiarities that make them splitting into 4 clusters rather than being clustered together, whereas the transients of cluster 5 (C^1_5) greatly differ from the others (outliers) [30].

Fig. 4 shows the three clusters C^2_1, C^2_2 and C^2_3 of $j = 2$ base clustering (circle, square and diamond markers, respectively, whereas the fourth cluster C^2_{NaN} is composed by transients with missing operational and environmental signal values and, thus, cannot be plotted). One can easily recognize that the shut-down transients belonging to C^2_3 (diamonds) are influenced by "High" temperature values at the turbine inlet and "High" vacuum values compared to C^2_1 and C^2_2.

The optimum number of clusters in the final consensus clustering is selected according to the Silhouette values for different numbers of clusters $C_{Candidate}$ that span in the interval [4,20], where the lower bound (4) is the minimum between C^1_{opt} and C^2_{opt}, and the upper bound (20) is the number of the largest combination of the two base clusters (i.e., 4×5) [32].

The optimum number of clusters in the final consensus clustering is found to be $P^*_{FF1} = 10$, at which the Silhouette measure is maximized (star in Fig. 5 (left)). Fig. 5 (left) shows, indeed, that the Silhouette values for small and large numbers of $C_{Candidate}$ are worse than for $P^*_{FF1} = 10$ (i.e., 0.8842), due to the dissimilarity of the data (inappropriately) assigned to the same clusters. The individual Silhouette values of each consensus cluster are shown in Fig. 5 (right): the Silhouette values for most of the clusters are equal to 1, which indicates the goodness of the obtained clusters in terms of separation and compactness ([34]; see Appendix C).

The obtained consensus results for $j = 1$ and $j = 2$ base clusterings are reported in Table 3.

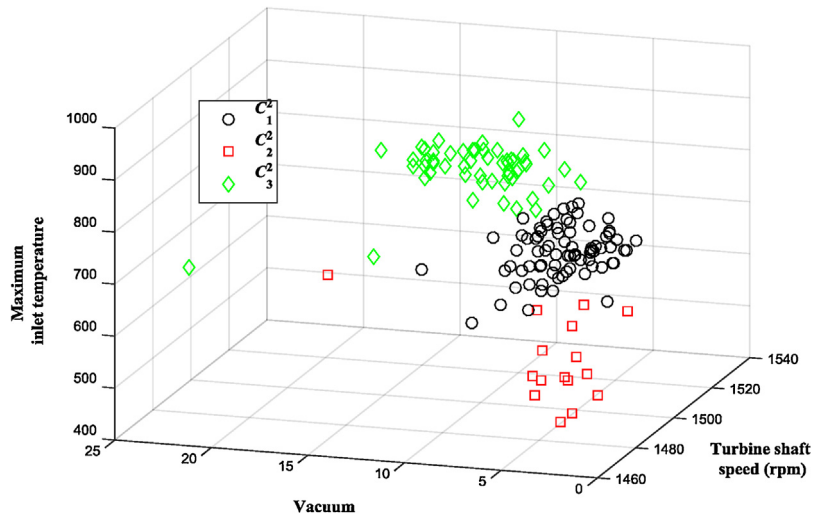

Fig. 4. Plot of the values of the clusters obtained by $j = 2$ base clustering C^2_1, C^2_2 and C^2_3.

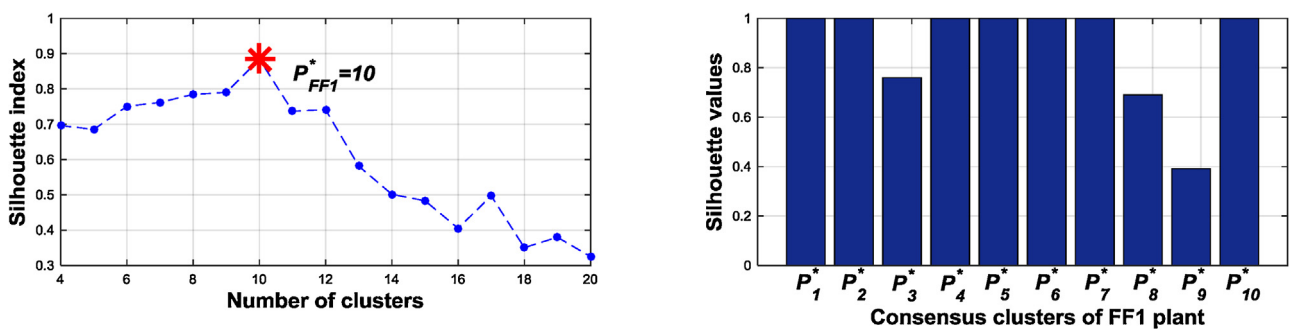

Fig. 5. Silhouette values vs. cluster numbers (left) and the individual Silhouette values of each consensus cluster obtained for the optimum number of clusters (right).

In fact, looking at the final consensus clustering one can recognize that transients of each cluster obtained by the $j = 1$ base clustering are influenced by different environmental and operational conditions that are obtained by the $j = 2$ base clustering. For example, the transients belonging to C^1_2 of the $j = 1$ base clustering have been splitted into three different final consensus clusters (P^*_5, P^*_6, and P^*_{10}) each one is due to a different enviornmental and operational condition (C^2_2, C^2_3, and C^2_1) as recognized by the $j = 2$ base clustering.

3.2. Step 2: reconciliating the most similar consensus clusters obtained individually for the two different plants

The available information of the other turbine (i.e., EE1) is used to update the previous obtained consensus clusters of FF1.

To avoid catastrophic forgetting and the need to retrain the diagnostic tool (as we shall see in Section 4), the proposed framework suggests a reconciliation approach that aims at learning the novel information content of the new EE1 transients without forgetting the previously acquired knowledge of the $P^*_{FF1} = 10$ consensus clusters of FF1 transients. To this aim, the $N_{EE1} = 116$ transients have been partitioned into $P^*_{EE1} = 16$ turbine behaviors explained by different environmental and operational conditions, by resorting to unsupervised ensemble clustering. Table 4 reports the number of base clusters along with the optimum number of clusters obtained in the final consensus clustering $P^*_{EE1} = 16$.

Fig. 6 shows the evolution of vibration signal 2 of the $j = 1$ base clustering (4 clusters) and the corresponding turbine speed values. One can easily recognize that, on one side, the functional behaviors

Table 4
Number of base clusters along with the optimum number of clusters in the final ensemble clustering of EE1 NPP.

j = 1 Base clustering	j = 2 Base clustering	*Consensus clustering*
$C^1_{opt} = 5$	$C^2_{opt} = 4$	$P^*_{EE1} = 16$

of transients belonging to clusters 1–3 (C^1_1, C^1_2 and C^1_3)are similar but with some peculiarities that make them splitting into 3 clusters rather than being clustered together, whereas the outliers (C^1_4) are different as they have a very dispersed and wide peak value around 700 rpm. It is worth mentioning that the fifth cluster C^1_{NaN} is composed by transients containing non-physical values of the vibration signals and, thus, cannot be plotted.

Fig. 7 shows the three clusters C^2_1, C^2_2 and C^2_3 of $j = 2$ base clustering (circle, square and diamond markers, respectively, whereas the fourth cluster C^2_{NaN} is composed by transients with missing operational and environmental signal values and, thus, cannot be plotted). One can easily recognize that the shut-down transients belonging to C^2_1 (circles) are influenced by "High" temperature values at the turbine inlet and "High" vacuum values compared to C^2_2 and C^2_3.

The optimum number of clusters in the final consensus clustering is selected according to the Silhouette values for different numbers of clusters $C_{Candidate}$ that span in the interval [4,20], where the lower bound (4) is the minimum between C^1_{opt} and C^2_{opt}, and the upper bound (20) is the number of the largest combination of the two base clusters (i.e., 4×5) [33].

The optimum number of clusters in the final consensus clustering is found to be

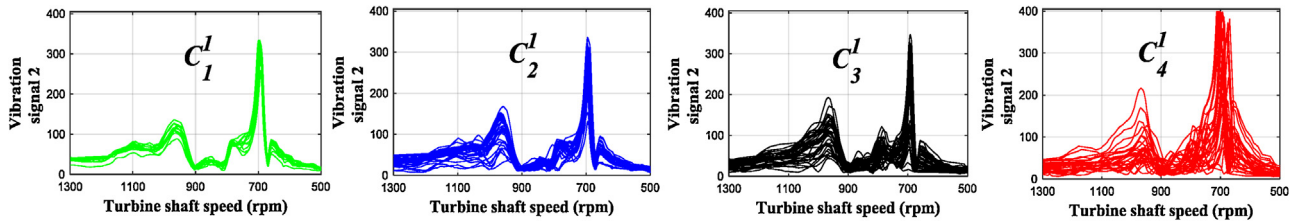

Fig. 6. Evolution of vibration signal 2 of the 4 clusters (C_1^1, C_2^1, C_3^1 and C_4^1) of the $j = 1$ base clustering and the corresponding turbine shaft speed values of EE1 *NPP* turbine.

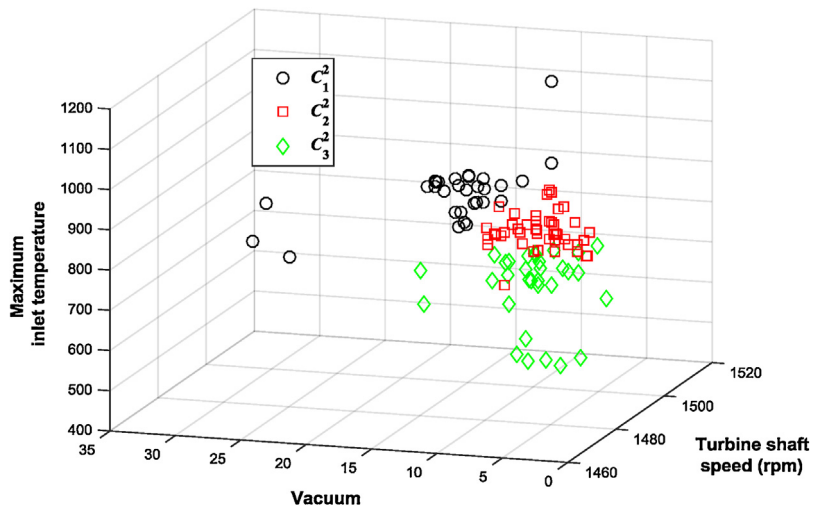

Fig. 7. Plot of the values of the clusters obtained by $j = 2$ base clustering C_1^2, C_2^2 and C_3^2.

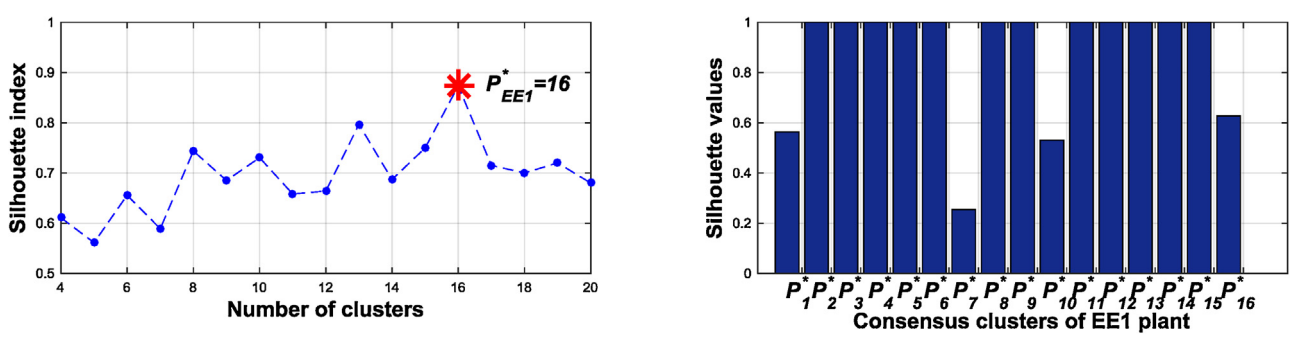

Fig. 8. Silhouette values vs. cluster numbers (left) and the individual Silhouette values of each consensus cluster obtained for the optimum number of clusters (right).

$P^*_{EE1} = 16$, at which the Silhouette measure is maximized (star in Fig. 8 (left)). Fig. 8 (left) shows, indeed, that the Silhouette values for small and large numbers of $C_{Candidate}$ are worse than for $P^*_{EE1} = 16$ (i.e., 0.8734), due to the dissimilarity of the data (inappropriately) assigned to the same clusters ([34]; see Appendix C). The individual Silhouette values of each consensus cluster are shown in Fig. 8 (right).

The obtained consensus results for $j = 1$ and $j = 2$ base clusterings are reported in Table 5.

To identify the consensus clusters among P^*_{FF1} and P^*_{EE1} with similar vibrational behaviors (i.e., hereafter called the best matching clusters), a preliminary visual analysis has been done for identifying the most similar vibration signals of the transients composed by each exhaustive pair of the consensus clusters P^*_{FF1} and P^*_{EE1}: the clusters with similar vibrational behaviors have been selected as best matching clusters (reported in Table 6).

For clarification purposes, Fig. 9 shows the seven vibration signals of the transients belonging to the identified best matching cluster # 1. Looking to the curves, one can recognize that the transients have, indeed, similar vibrational behaviors. In fact, the differences among the transients of each best matching cluster are related to the magnitude of the peaks, e.g., looking to the figure, the magnitude of the peak around the turbine shaft speed of 800 rpm of vibration signal 3 is larger for P^*_9 transients than for the transients of P^*_{11}, and/or to the delay in the transients occurrences. However, their functional behaviors are similar and, hence, they have been selected as the best matching clusters between the plants. This suggests us that, practically, one can identify the best matching clusters by resorting to functional similarity methods [48,49] for quantifying the extent of similarity of clusters obtained for each plant independently (this will be the focus of future research work).

Furthermore, Fig. 10 (right) shows the environmental and operational conditions mean values ($j = 2$ base clustering) (dots) of the transients composed by the best matching clusters of FF1 (P^*_9, P^*_8, P^*_2) and EE1 (P^*_{11}, P^*_5, P^*_7). Looking at Fig. 10 (right), one can notice that:

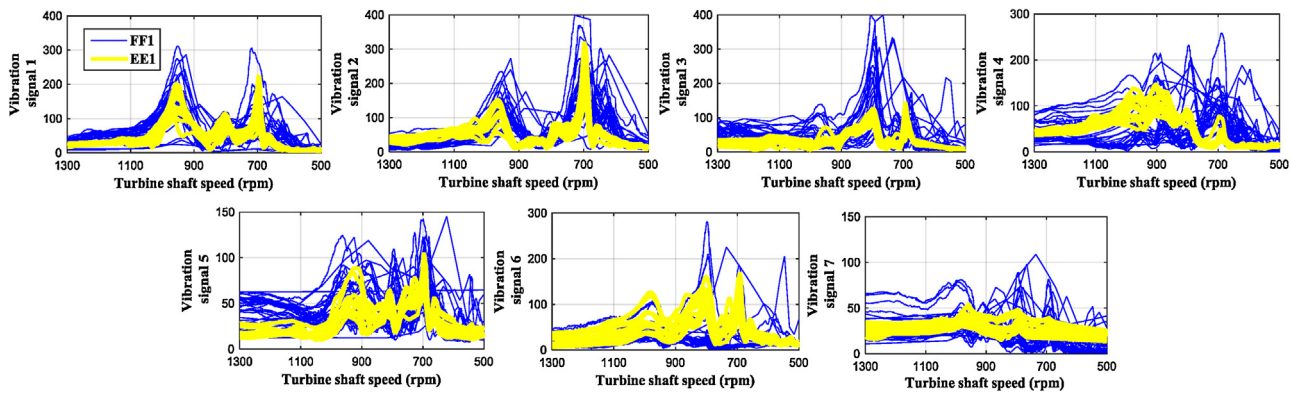

Fig. 9. Evolution of the 7 vibration signals of the best matching clusters P_9^* (FF1) with P_{11}^* (EE1).

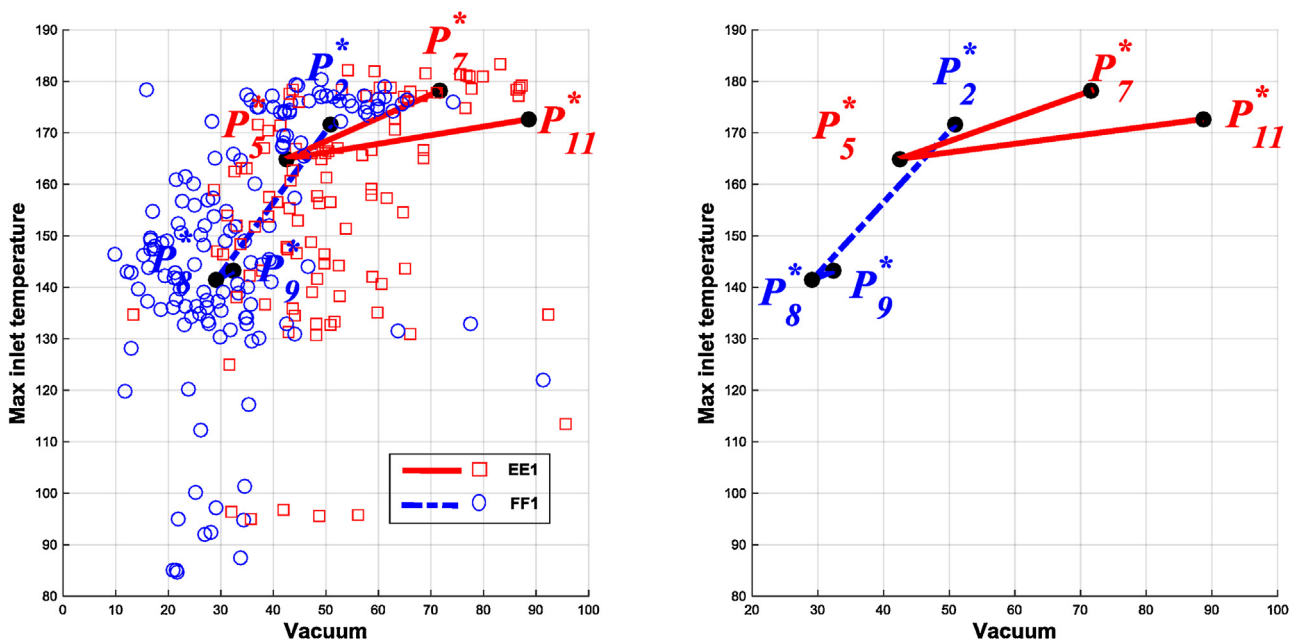

Fig. 10. Plot of the $j=2$ base clustering mean values (dots) of the 3 best matching clusters (right) along with $j=2$ base clustering values of all transients of the two *NPPs* turbines (left).

Table 5
Consensus results for $j=1$ and $j=2$ base clusterings.

$j=1$ Base clustering	$j=2$ Base clustering	*Consensus clustering*
C_1^1	C_1^2	P_7^*
	C_2^2	P_5^*
	C_{NaN}^2	P_{15}^*
C_2^1	C_1^2	P_{11}^*
	C_2^2	P_3^*
	C_3^2	P_{13}^*
C_3^1	C_1^2	P_1^*
	C_2^2	P_{12}^*
	C_3^2	P_{14}^*
C_4^1	C_1^2	P_6^*
	C_2^2	P_2^*
	C_3^2	P_4^*
	C_{NaN}^2	P_{16}^*
C_{NaN}^1	C_1^2	P_9^*
	C_2^2	P_8^*
	C_3^2	P_{10}^*

Table 6
Best matching consensus clusters of FF1 and EE1 NPPs turbines.

	Best Matching cluster 1	*Best Matching cluster 2*	*Best Matching cluster 3*
FF1	P_9^*	P_8^*	P_2^*
EE1	P_{11}^*	P_5^*	P_7^*

1) The evolution of the turbines behaviors (i.e., dashed line for FF1 and solid line for EE1) through the three successive behaviors of the best matching clusters is similar for the two plants due to the similar changes of the environmental and operational conditions that influence the turbines behavior. For example, the evolution of the turbine behavior of the FF1 *NPP* from P_9^* to P_8^* is similar to the evolution of the turbine behavior of the EE1 *NPP* from P_{11}^* to P_5^*: both are caused by a decrease of the vacuum value and of the maximum inlet temperature value. Similarly, the evolution of the turbine behavior of the FF1 *NPP* from P_8^* to P_2^* is similar to the evolution of the turbine behavior of the EE1 *NPP* from P_5^* to P_7^*: both are caused by an increase of the vacuum value and of the maximum inlet temperature value.
2) The dashed and solid lines are displaced from each other in the environmental and operational conditions space. This can be justified by the fact that the two turbines are influenced by environmental and operational conditions displaced from each other as shown in Fig. 10 (left):

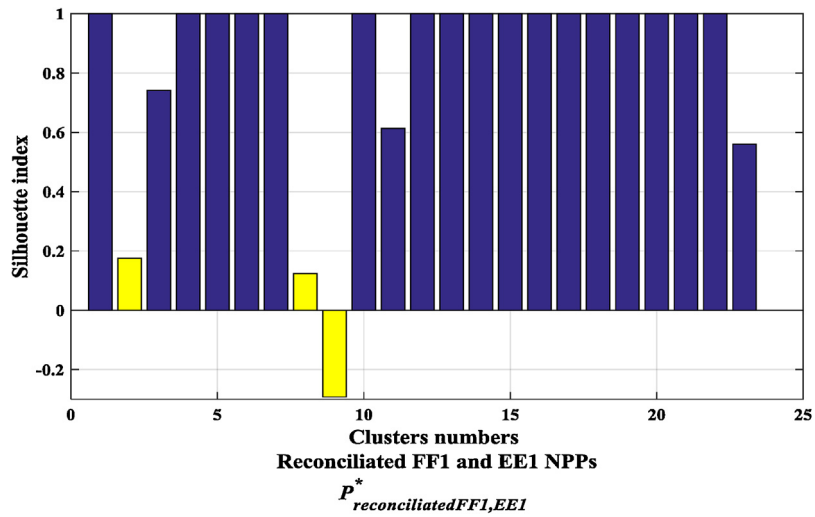

Fig. 11. Silhouette values for the obtained consensus clusters of the reconciliated final ensemble clustering.

- the shut-down transients belonging to the EE1 plant (squares) are influenced, on average, by "high" vacuum values and "high" temperatures values at the turbine inlet,
- the shut-down transients belonging to the FF1 plant (circles) are influenced, on average, by "low" vacuum values and "low" temperatures values at the turbine inlet.

In addition to that, the reconciliation strategy leads to the reduction of the consensus clusters to 23 instead of the 26 consensus clusters individually obtained for the two plants:

1) 7 clusters contain transients with different vibrational behaviors caused by different environmental and operational conditions, but uniquely occurred in FF1 *NPP*,
2) 13 clusters contain transients with different vibrational behaviors caused by different environmental and operational conditions, but uniquely occurred in EE1 *NPP*,
3) 3 clusters contain transients with different vibrational behaviors caused by different environmental and operational conditions, but similarly occurred in both plants.

The goodness of the final obtained clusters is verified with respect to the cluster separation and compactness by resorting to the Silhouette validity index ([34]; see Appendix C)), *C-index* [35] and *DB* criterion [36]. Fig. 11 shows the individual Silhouette values of the 23 consensus clusters:

1) Most of the clusters that belong either only to FF1 or to EE1 (dark shade of color) have Silhouette values equal to 1, which indicates that these clusters are composed by very similar transients,
2) The Silhouette values of the three reconciliated clusters (light shade of color) have low Silhouette values compared to the others. This can be explained by the fact that either the transients from the two plants are poorly-matched and/or the consensus clusters independently obtained for the individual two plants are not originally composed by transients similar to each other (low Silhouette values as shown in Figs. 5 and 8 for FF1 and EE1, respectively).

In fact, the consensus cluster P^*_2 of FF1 has a Silhouette value of 1 which indicates that the transients composed by the cluster are well-matched to each other within the cluster and poorly-matched to other transients in the other clusters with respect to both $j = 1$ and $j = 2$ base clusterings (Fig. 5), whereas the Silhouette value of the consensus cluster P^*_7 of EE1 is ~0.2, which indicates that the transients composed by the cluster are poorly-matched to each other within the cluster but with some similarities to those in the other clusters with respect to both $j = 1$ and $j = 2$ base clusterings (Fig. 8).

Table 7
Average C-index and DB values of the final consensus clusters, compared to FF1 and EE1 consensus clusters.

	C-index	*DB*
P^*_{FF1}	0.0274	0.4973
$\boldsymbol{P^*_{EE1}}$	0.0281	0.4247
P^*_{final}	0.0304	0.6162

1) The average Silhouette value of the overall 23 consensus clusters is equal to 0.8227. The value is, indeed, representative of the goodness of the final obtained clusters.

For completeness, Table 7 reports the average *C-index* and *DB* values of the P^*_{final} consensus clusters compared to the individual P^*_{FF1} and P^*_{EE1} consensus clusters of FF1 and EE1 *NPPs* turbines, respectively. One can notice that:

- the small *C-index* and *DB* values (i.e., close to 0) confirm (in support to 3) above) the goodness of the obtained P^*_{FF1}, P^*_{EE1} and P^*_{final} consensus clusters, in terms of their separation and compactness;
- the *C-index* and *DB* values of the P^*_{final} consensus clusters are slightly higher (but still good) than those of FF1 and EE1 *NPPs* turbines, as discussed in 2) above.

It is worth mentioning that the unsupervised ensemble clustering (Step 1) has been performed with a Matlab code that has been in-house developed at the Laboratorio di Analisi di Segnale e Analisi di Rischio (LASAR, Laboratory of Signal and Risk Analysis of the Department of Energy of the Politecnico di Milano (www.lasar.polimi.it)); the computational time needed to run the code on an Intel Core i5 with data taken from the $P = 2$ *NPPs* turbines with 149 and 116 shut-down transients, respectively, is equal to 5 min. Since the proposed reconciliation approach (Step 2) matches (both by visual analysis and similarity matching) the results previously obtained by ensemble clustering performed (independently) on the N_p and N_{p+1} transients, the proposed framework is expected to be capable of integrating consensus clusters of individual plants with

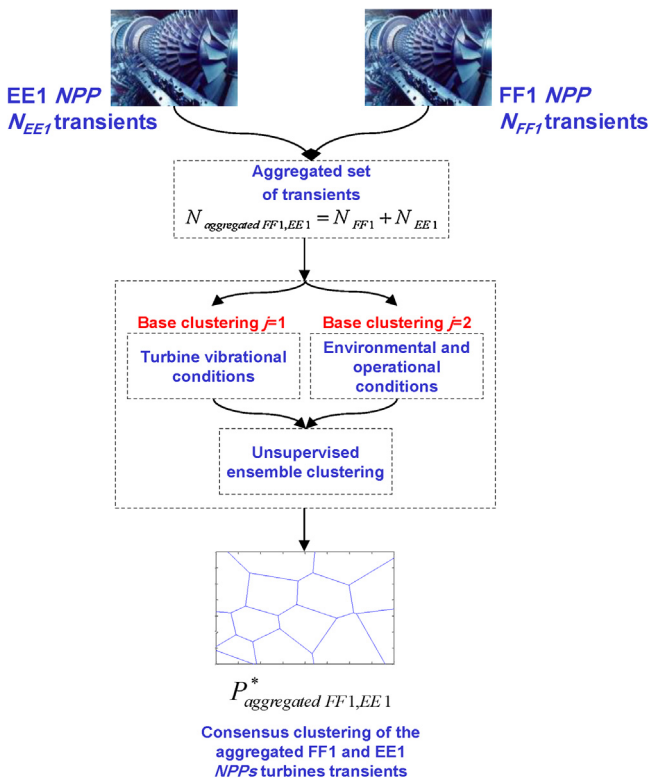

Fig. 12. Clustering of the aggregated shut-down transients of FF1 and EE1 NPPs turbines by the unsupervised ensemble clustering approach.

low computational efforts, also when dealing with even larger fleet dimensions.

However, to better verify the performance of the proposed framework, three alternative approaches are applied to the case study and their results are compared with those obtained by the proposed framework.

4. Comparison with other approaches

For comparing the results of the proposed framework for reconciliating different final ensemble clusters of different turbines, three other approaches are used, alternatively, to reconciliate the consensus clusters of the FF1 *NPP* turbine on the basis of the received information from the EE1 *NPP* turbine that are: 1) clustering of the aggregated shut-down transients of the FF1 and EE1 *NPPs* turbines by the unsupervised ensemble clustering approach, 2) the inclusion of the EE1 transients into FF1 ensemble clustering by resorting to Fuzzy similarity measure [37–39] and 3) the classification of the EE1 transients by a supervised classifier, such as a Fuzzy *K*-Nearest Neighbours algorithm (*FKNN*) [40–42] trained on FF1 consensus clustering.

4.1. Clustering of the aggregated shut-down transients of the FF1 and EE1 NPPs turbines by the unsupervised ensemble clustering approach

This approach aims at aggregating the transients of the available $p+1$-th component, $p = 1, .., P$, (resulting in $N_{aggregated1,2,\ldots,p+1} = N_1 + N_2 + \ldots + N_{p+1}$), and clustering them into $P^*_{aggregated1,2,\ldots,p+1}$ groups by resorting to the ensemble clustering approach of Section 2, Step 1 (see Appendix A), without any reconciliation (Step 2).

Fig. 12 shows its application to the two available dataset of the FF1 and EE1 plants: the approach starts by aggregating the transients of the two plants (resulting $N_{aggregatedFF1,EE1} = 149 + 116 = 265$) and, then, clustering them into $P^*_{aggregatedFF1,EE1}$ groups by resorting to the ensemble clustering approach of Section 2, Step 1.

Table 8
Number of base clusters along with the optimum number of clusters in the final ensemble clustering of the aggregated FF1 and EE1 NPPs shut-down transients.

$j=1$ Base clustering	$j=2$ Base clustering	*Consensus clustering*
$C^1_{opt} = 5$	$C^2_{opt} = 4$	$P^*_{aggregatedFF1,EE1} = 12$

Table 9
Consensus clusters of the aggregated shut-down transients of the two plants.

Consensus of FF1	P^*_3	P^*_5	P^*_6	P^*_7		
Consensus of EE1	P^*_8	P^*_{11}				
Common consensus	P^*_1	P^*_2	P^*_4	P^*_9	P^*_{10}	P^*_{12}

Table 8 reports the number of base clusters along with the optimum number of clusters obtained in the final consensus clustering that turns out to be $P^*_{aggregatedFF1,EE1} = 12$. The optimum number has been selected according to the Silhouette values for different numbers of clusters $C_{Candidate}$ that span in the interval [4,20], where the lower bound (4) is the minimum between C^1_{opt} and C^2_{opt} and the upper bound (20) is the number of the largest combination of the two base clusters (i.e., 4×5) (see star in Fig. 13 (left) that else shows that the Silhouette values for small and large numbers of $C_{Candidate}$ are worse than for $P^*_{aggregatedFF1,EE1} = 12$ (i.e., 0.8258), due to the dissimilarity of the data (inappropriately) assigned to the same clusters).

The individual Silhouette values of each of the obtained consensus clusters of the

$P^*_{aggregatedFF1,EE1} = 12$ are shown in Fig. 13 (right): the Silhouette values for most of the clusters are equal to 1, which indicates the goodness of the obtained clusters in terms of separation and compactness [34].

In this regard, 12 different groups of shut-down transients, representative of different behaviors explained by different environmental and operational conditions are obtained. It is worth mentioning that some consensus clusters are mainly composed by transients occurring independently in each plant (unique consensus clusters for each plant), whereas some others are composed by a combination of the transients occurring in the two plants (common consensus clusters between the plants), as reported in Table 9.

Moreover, it is worth mentioning that two clusters P^*_9 and P^*_{10} of the final consensus clustering aggregate the outliers which belong to C^1_5 and C^1_4 of the $j = 1$ base clustering and occur in the FF1 and EE1 plants, respectively. However, despite the fact that the overall Silhouette value (i.e., 0.8258) and also the *C-index* value (i.e., 0.0194) and the *DB* value (i.e., 0.3262) indicate the goodness of the obtained clusters, this approach:

1) entails discarding the existing $P^*_{FF1} = 10$ consensus clusters (representative of the FF1 plant turbine behaviors under different environmental and operational conditions) and computation-costly retraining the diagnostic tool with all data that have been accumulated thus far (i.e., $N_{aggregatedFF1,EE1} = 149 + 116 = 265$). This approach would result in a catastrophic forgetting of the acquired information contained in the $P^*_{FF1} = 10$ consensus clusters [50,51], and hence, the detailed analysis of each plant will be discarded once the whole transients are aggregated together. For this reason, this approach cannot be used to predict the health state of new incoming *NPP* turbines,
2) is considered infeasible for real diagnostic systems due to the computational efforts required for retraining on a large number of transients from a large number of plants [9]. In fact, the

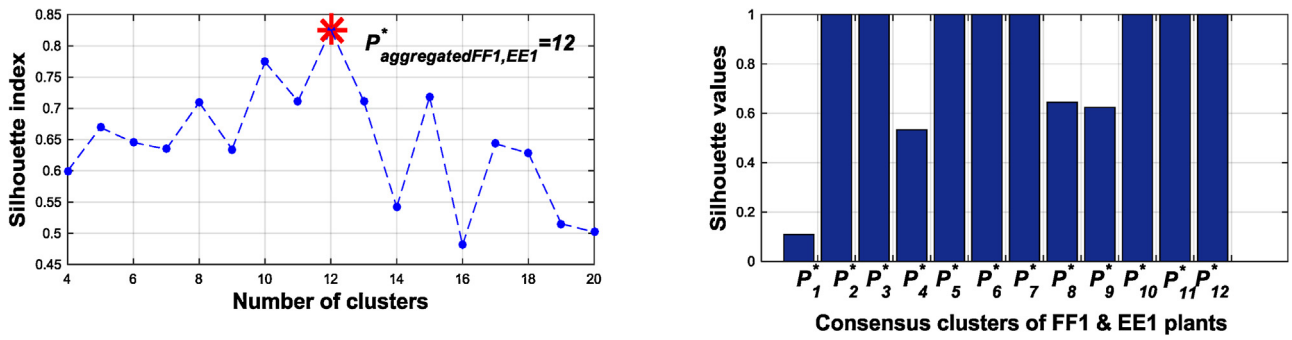

Fig. 13. Silhouette values vs. cluster numbers (left) and the individual Silhouette values of each consensus cluster obtained for the optimum number of clusters (right).

Inputs: P^*_{FF1} consensus clusters of the FF1 *NPP* turbine (Section 3.1) and the N_{EE1} new coming transients of EE1 *NPP* turbine
Outputs: Allocation of EE1 transients to one of FF1 consensus clusters with a fuzzy similarity measure

Step 1: Quantification of the similarity of a new coming transient of the EE1 *NPP* with respect to the transients belonging to each P^*_{FF1} consensus clusters previously obtained for the individual FF1 *NPP*

for *e*=1:105 % Number of transients of the EE1 *NPP*
  for *m*=1:10 % Number of consensus clusters of the FF1 *NPP*
    for *f*=1:length($P^*_{FF1}(m)$) % Number of transients belonging to each consensus clusters of the FF1 *NPP*
      • Quantify the similarity of each *e*-th transient with respect to the transients belonging to each *m*-th consensus cluster, following Eqs. (1) and (2).
    end
    • Compute the corresponding mean similarity value, following Eq. (3).
  end
end

Step 2: Allocate the new coming transient to any of the P^*_{FF1} consensus clusters

for *e*=1:105 % Number of transients of the EE1 *NPP*
  • Allocate the new coming *e*-th transient to the consensus cluster for which the similarity is the largest.
end

Fig. 14. The pseudo-code of the procedural steps of the alternative approach 2.

application of the unsupervised ensemble clustering approach to the aggregated set of transients of the two *NPPs* turbines (i.e., 256 transients) requires 12.6 min, which is more than double of the 5 min required by the reconciliation framework proposed in Section 2.

4.2. *Inclusion of a new transient occurring in the EE1* NPP *turbine into the ensemble clustering of FF1* NPP *turbine*

Another approach consists in the inclusion of the shut-down transients of a new $p+1$-th component, into the reference consensus clusters obtained from the previous available components, i.e., $1, \ldots, p$. The idea is that once the $P^*_{1,\ldots,p}$ ensemble clustering of the p components is obtained, the fuzzy similarity measure [37–39] is used to verify its capability to accommodate the new N_{p+1} transients of the $p+1$-th component by assigning the transients to the reference consensus clusters for which the obtained similarity is the largest. In this way, we can avoid extra costs for training the diagnostic tool and/or overtraining it (Section 4.1).

The procedural developments of the approach with respect to the available two *NPPs* turbines are summarized in the following two steps and reported in the pseudo-code of Fig. 14. It is worth mentioning that a group of 11 transients containing non-physical values of the vibration signals of $j=1$ base clustering has been excluded from the analysis and, hence, $N_{EE1} = 116 - 11 = 105$.

Step 1: *Quantification of the similarity of a new coming transient of the EE1 NPP with respect to the transients belonging to each P^*_{FF1} consensus clusters previously obtained for the individual FF1 NPP.* To this aim, a fuzzy similarity measure is used to determine the degree of closeness of each e-th transient, $e = 1, \ldots, 105$ occurring in EE1 plant with each f-th transient occurring in FF1 plant and allocated to the m-th consensus cluster, $f = 1, \ldots, length\left(P^*_{FF1}(m)\right)$, $m = 1, \ldots, 10$, with reference to the pointwise difference between the values of the matrices $\overset{=e}{Y}$ [7800] and $\overset{=f}{Y}$ [7800], where $y^{e/f}_{zt}$ is the t-th vibrational measurement, $t = 1, \ldots, 800$, of the z-th vibrational signal, $z = 1, \ldots, 7$, of matrix $\overset{=e}{Y}$ and $\overset{=f}{Y}$, respectively [30]. The pointwise difference δ^m_{ef} between the 7*800=5600 values of transient $\overset{=e}{Y}$ and transient $\overset{=f}{Y}$ of the m-th consensus cluster is defined by Eq. (1):

$$\delta^m_{ef} = \sqrt{\sum_{z=1}^{7}\sum_{t=1}^{800}\left(y^e_{zt} - y^f_{zt}\right)^2} \tag{1}$$

The pointwise difference of the two transients is then evaluated with reference to an "approximately zero" fuzzy set (*FS*) specified by a function which maps δ^m_{ef} into a value μ^m_{ef} of membership to the condition of "approximately zero": values of μ^m_{ef} close to 0 indicate that the signal evolutions in the two transients e and f are very different, whereas values close to 1 indicate high similarity. In this work, the bell-shaped function shown in Eq. (2) is used with the optimum value of the bell-shaped function parameter σ at $\sigma = 2.92$:

$$\mu^m_{ef} = e^{\frac{\left(\delta^m_{ef}\right)^2}{\sigma^2}} \tag{2}$$

Then, the mean similarity value of each i-th transient of the EE1 plant with all the transients of each m-th consensus cluster of FF1 plant is quantified as shown in Eq. (3).

$$S^m_e = mean_f\left(\mu^m_{ef}\right), f = 1, \ldots, length\left(P^*_{FF1}(m)\right), m = 1, \ldots, 10 \tag{3}$$

Step 2: *Allocate the new coming transient to any of the P^*_{FF1} consensus clusters.* In this Step, all of the transients of EE1 *NPP* turbine ($N_{EE1} = 116 - 11 = 105$) will be allocated to one of FF1 consensus clusters for which the mean similarity measure is the largest.

To verify if the consensus clusters fit the allocated transients of EE1 or not, one way can be by clustering the mean similarity values of EE1 transients into dissimilar groups, whose number is "a priori" unknown. The motivation of doing this is to split the similarity values of the transients' allocation into different groups with different categories of the similarity values and, then, select those with largest similarity values with respect to a fixed threshold.

In this regard, the optimum number of clusters C^* is selected according to the values of the Silhouette index [34] for different numbers of clusters $C_{Candidate}$ that span the interval [2,13]: the optimum number of clusters C^* is the value at which the Silhouette is maximized, i.e., $C^* = 2$ (star in Fig. 15 (left)).

The clustering results are shown in Fig. 15 (right): one can consider a fixed threshold value of 0.12 for which the transients with similarity values larger than the threshold (i.e., transients of cluster 2–triangles markers in Fig. 15 (right), respectively) are considered well allocated to FF1 consensus clusters with high confidence level, whereas the transients with similarities lower than the threshold (i.e., transients of cluster 1– squares markers in Fig. 15 (right),

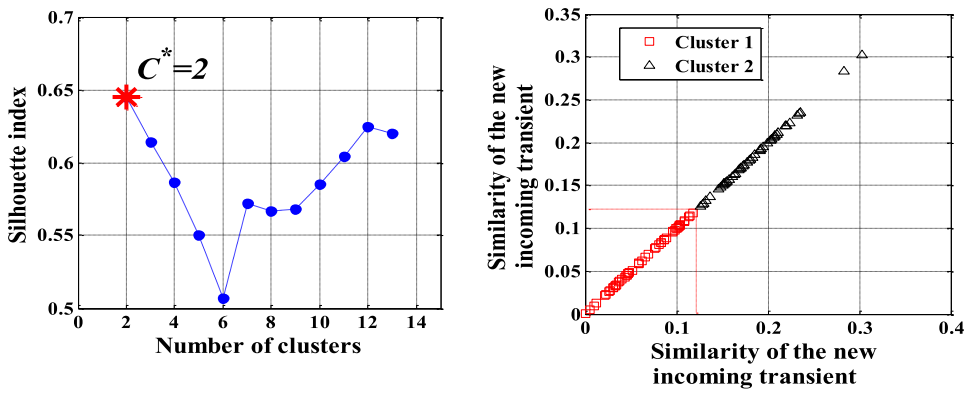

Fig. 15. Silhouette values vs. cluster numbers (left) and clustering results of the mean similarity values of the new coming transients (right).

Table 10
Average Silhouette, C-index and DB values of the updated FF1 consensus clusters after the allocations of EE1 transients by the fuzzy similarity measure.

Silhouette	*C-index*	*DB*
0.2594	*0.1214*	*1.7601*

respectively) are considered allocated to FF1 consensus clusters with low confidence levels. In this regard, one can conclude that FF1 consensus clustering cannot accommodate the whole transients of the EE1 *NPP* turbine.

For example, Fig. 16 shows the mean similarity values of transient $e = 23$ (transient from cluster 2) (Fig. 16 (left)) and $e = 88$ (transient from cluster 1) (Fig. 16 (right)) of the EE1 turbine, respectively with respect to the $P^*_{FF1} = 10$ consensus clusters of FF1 plant. It is easy noticing that the largest similarity values of the two transients occur with the transients of P^*_5 of FF1 (light shade of color).

For the sake of clarity, Fig. 17 shows the transients of P^*_5 of FF1 (solid) along with the allocated EE1 transients $e = 23$ and $e = 88$ (dashed) for three vibration signals 1, 2 and 3. It is worth noticing that the transient $e = 23$ (Fig. 17 (left)) seems to be well allocated to P^*_5 of FF1, whereas the transient $e = 88$ (Fig. 17 (right)) does not (allocated with low confidence level); for example, the transient has no peak for vibration signal 1 around the turbine shaft speed of 920 rpm.

To quantify the influence of EE1 transients' allocations on the goodness of the updated FF1 consensus clusters, Table 10 reports the average Silhouette, *C-index* and *DB* values of the overall updated FF1 consensus clusters. One can easily recognize that the quality of the final clusters, in terms of cluster separation and compactness, is much worse than for the proposed approach (Table 7 and Fig. 11), as shown by the small Silhouette value and large *C-index* and *DB* values.

For completeness, the application of this approach to the available dataset requires 4.5 min, which is very much in line with the computational efforts required by the reconciliation framework proposed.

4.3. Classifying the shut-down transients of the EE1 NPP turbine by a classifier trained on FF1 consensus clusters

In this approach, the objective is to use the labelled transients of the $P^*_{1,\ldots,p}$ consensus clusters previously obtained for the p components, i.e., $p = 1, \ldots, P$, as reference trajectories for training a supervised classifier. The resulting classifier is, then, used to classify the new transients of $p + 1$-th component into the appropriate consensus clusters of the p plants with associated membership values.

Several classification algorithms have been proposed and used in practice, like Support Vector Machines (*SVM*) [52], Naïve Bayes classifier [53], Decision trees [54], Discriminant analysis [55], Classification and Regression Tree (*CART*) [56,57] and Fuzzy *K*-Nearest Neighbours (*FKNN*) [40–42]. In this work, we resort to the Fuzzy *K*-Nearest Neighbors (*FKNN*) algorithm, because *FKNN* is simple, requires less computation time during the training phase and is one of the most used [42].

The procedural steps for the application of this approach to the available two *NPPs* turbines are given in the pseudo-code of Fig. 18. The approach entails the following two steps:

Step 1: *Training the Fuzzy K-Nearest Neighbours with FF1 data.* To this aim, the training dataset $\overline{\overline{X}}_{FF1}$ is constructed on the basis of the data available from FF1 *NPP* turbine summarized in the $P^*_{FF1} = 10$ consensus clusters, as we have seen in Section 2. Therefore, the training dataset $\overline{\overline{X}}_{FF1}$ is composed by:

i) data representative of the turbine condition ($j = 1$), and of the environmental and operational conditions that can influence the vibrations ($j = 2$), and
ii) knowledge represented by the 10 labels obtained by the application of the unsupervised ensemble clustering algorithm on the information available from the FF1 *NPP* turbine [32].

Step 2: *Classifying new coming transients from the EE1 NPP.* Once the *FKNN* classifier has been trained on FF1 data, the trained *FKNN* is used to classify the N_{EE1} transients of the EE1 turbine to the consensus clusters with associated membership values. The obtained results are stored in matrix $\overline{\overline{Y}}_{EE1}$ with a size of N_{EE1}x2.

The basic idea of *FKNN* classifier is to determine the unknown consensus cluster of a new coming transient of the EE1 turbine by looking at the known consensus clusters of its neighbors. More specifically, the classification of a new coming e-th transient of the EE1 turbine, $e = 1, \ldots, N_{EE1}$, is done by assignment to the m-th consensus cluster, $m = 1, \ldots, P^*_{FF1}$, with a certain membership value μ^m_e, that is the highest among the other membership values to the other consensus clusters. The membership value is a function of the distances between the new coming e-th transient and its Kth nearest neighbors and their consensus clusters memberships, where the Kth nearest neighbors are the K transients of the training dataset that are closest to the new transient according to the Euclidean distance [42].

It is worth mentioning that the *FKNN* is trained on the training dataset $\overline{\overline{X}}_{FF1}$ via the 10-fold Cross-Validation (*CV*) analysis [58] and its classification performance is evaluated by resorting to external validity measures [59], e.g., false-positive rate and false-negative rate.

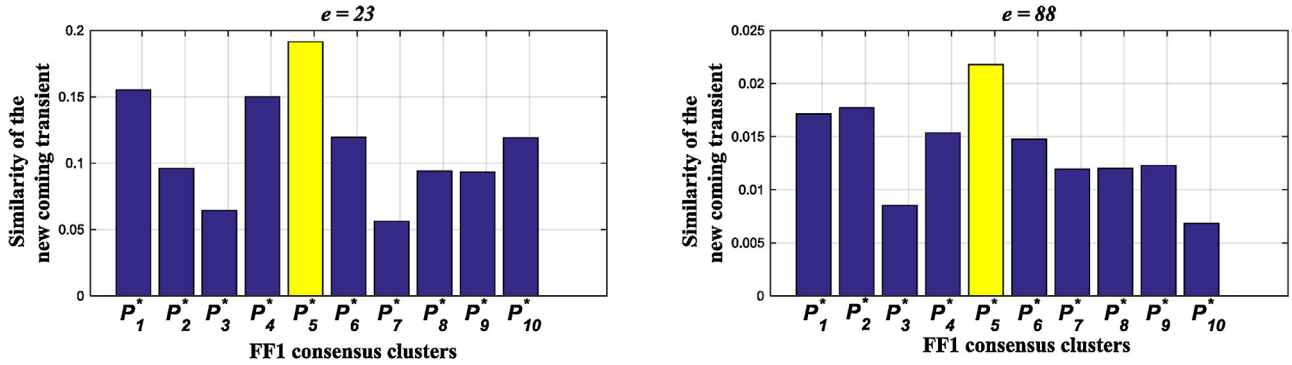

Fig. 16. Mean similarity values of transients $\boldsymbol{e} = 23$ (left) and $\boldsymbol{e} = 88$ (right) with respect to the $\boldsymbol{P}^*_{FF1} = 10$ consensus clusters of FF1 plant.

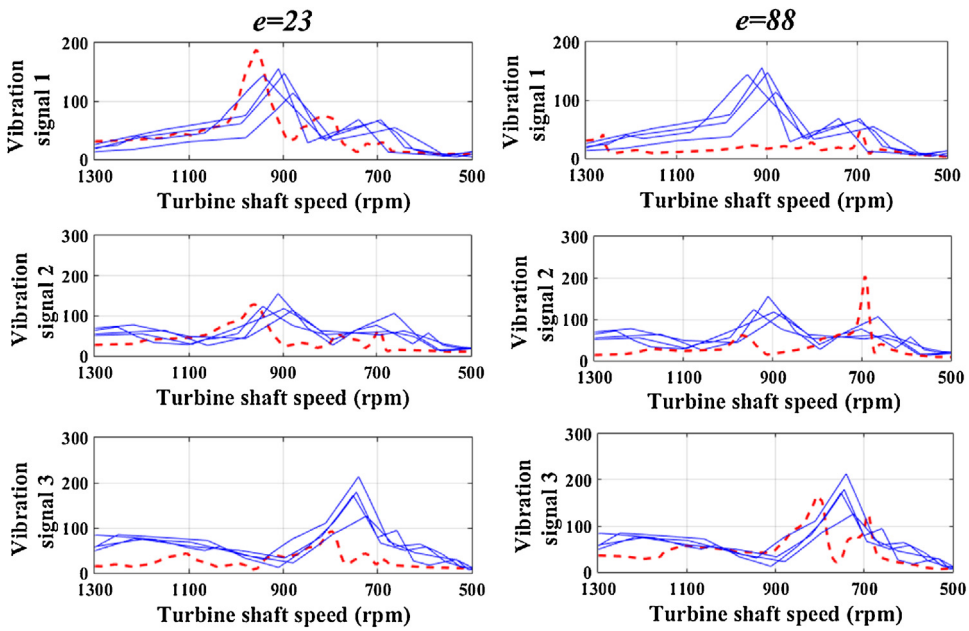

Fig. 17. Transients $e = 23$ (left) and $e = 88$ (right) with the transients of P^*_5 of FF1 NPP.

Inputs: $\bar{\bar{X}}_{FF1}$ training dataset of FF1 *NPP* turbine with the 10 labels obained in Section 3.1 and the EE1 transients which need to be allocated
Outputs: Allocation of EE1 transients to one of FF1 consensus clusters with a membership value

Step 1: Training the *FKNN* with $\bar{\bar{X}}_{FF1}$ training dataset
Step 2: Classifying new coming transients from the EE1 *NPP* turbine
for $e = 1: N_{EE1}$ % Number of new coming transients of the EE1 *NPP*
- Calculate the Euclidean distances between the *e*-th transient and the whole transients of FF1
- Identify the Kth nearest neighbor transients of the training dataset that are closest to the *e*-th coming transient (Kth smallest Euclidean distances)
- Calculate the membership values of the *e*-th transient to all of the FF1 consensus clusters as a function of its distances to the Kth nearest transients and their consensus clusters memberships
- Assign the *e*-th transient to the *m*-th consensus cluster, $m = 1, \ldots, P^*_{FF1}$, with a certain membership value μ^m_e, that is the highest among the other membership values of the *e*-th transient to the other consensus clusters.

end

Fig. 18. The pseudo-code of the procedural steps of the alternative approach 3.

Fig. 19 shows the largest membership values of the EE1 transients classified to FF1 consensus clusters using the optimum K value, $K^* = 7$: the optimum K value is identified via a sensitivity analysis performed for investigating the influence of different K values on the quality of the updated FF1 consensus clusters, in terms of clusters separation and compactness. The K value at which the clusters are well separated and compactness is selected to be the optimum value (refer to Appendix D for more details on the selection procedure of the K^*). It is worth mentioning that among $N_{EE1} = 116$ transients, the groups of 12 transients and 8 transients containing non-physical values of the signals used for $j = 1$ and $j = 2$ base clusterings, respectively have been excluded from the analysis ($N_{EE1} = 116 - 20 = 96$).

For evaluating the results, one can consider a fixed threshold value of 0.5: if the transient has a membership value larger than the threshold (i.e., 70 out of 96 transients), it is considered well allocated to FF1 consensus clusters with high confidence level (circles); on the contrary, if a transient has a membership value lower than the threshold (i.e., 26 out of 96 transients), then, it is considered allocated to FF1 consensus clusters with low confidence levels

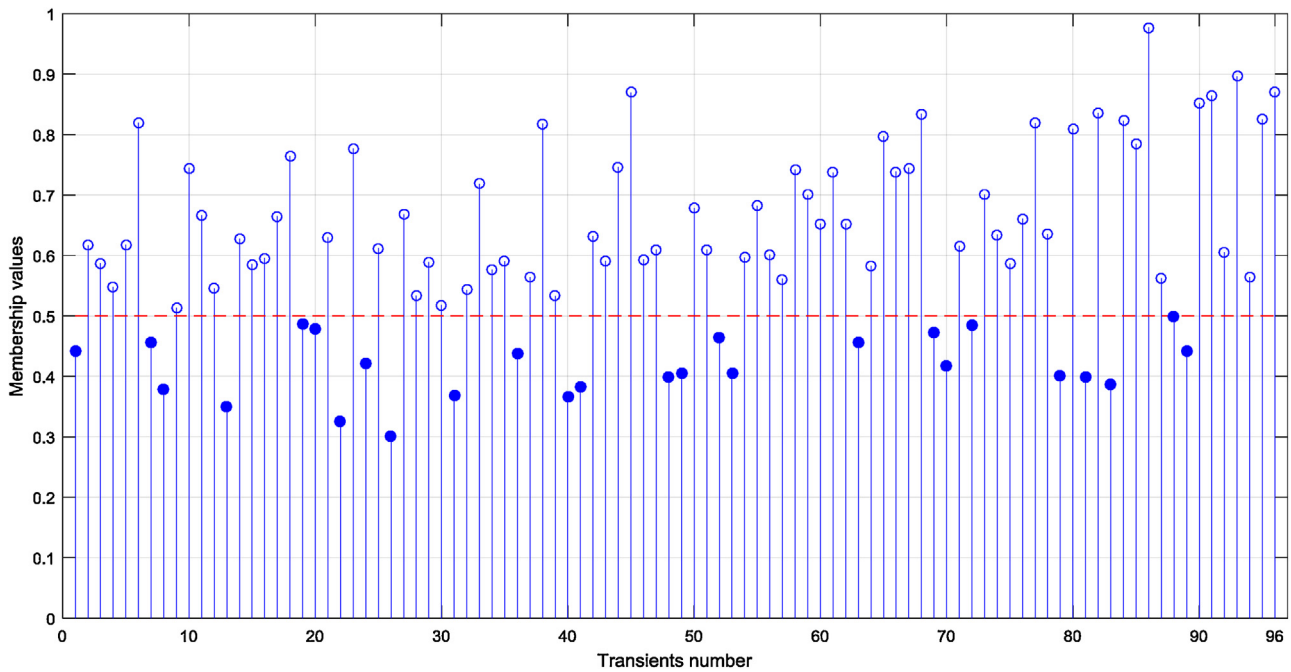

Fig. 19. Largest membership values for each transient of EE1 plant.

Table 11
Average Silhouette, C-index and DB values of the updated FF1 consensus clusters obtained by the three alternative approaches compared with the proposed approach, along with their computational efforts.

	Silhouette	*C-index*	*DB*	*Computational Efforts (min)*
Alternative approach 1	*0.8258*	*0.0194*	*0.3262*	*12.6*
Alternative approach 2	*0.2594*	*0.1214*	*1.7601*	*4.5*
Alternative approach 3	*−0.0230*	0.5735	*3.2235*	*4.7*
Proposed approach	*0.8227*	*0.0304*	*0.6162*	*5*

(dots). Table 11 summarizes the average Silhouette, *C-index* and *DB* values obtained by the three alternative approaches compared with the proposed approach, along with their computational efforts. One can notice that:

- the overall values of the Silhouette (i.e., −0.0230), *C-index* (i.e., 0.5735) and *DB* (i.e., 3.2235) of the updated FF1 consensus clusters obtained at $K^*$by the alternative approach 3 ensure that the quality of the obtained clusters, in terms of separation and compactness, is much worse than for the proposed approach (Fig. 11), as shown by the very small Silhouette value and very large *C-index* and *DB* values,
- the application of this approach to the available dataset requires a computational effort (i.e., 4.7 min) almost similar to the reconciliation framework proposed (i.e., 5 min).

As last remark, one might be wondering whether the distributive characteristics of the available training dataset $\overline{\overline{X}}_{FF1}$, i.e., proportions of number of transients in the available clusters, have an impact on the optimum K value and the quality of the final updated consensus clusters. This is investigated and its detailed analysis is reported in Appendix E, for completeness: the obtained results show that the $\overline{\overline{X}}_{FF1}$ is a low imbalanced dataset (i.e., the number of transients in the majority clusters is non-significantly larger than that in the minority clusters) and it has no influence on the identified optimum K value, K^*, and correspondingly, on the final conclusions drawn.

5. Conclusions and perspectives

In this paper, we have proposed a framework for incrementally learning the different clusterings independently obtained for *NPP* individual turbines of a fleet. The basic idea is to reconciliate the most similar clusters in the different *NPP* turbines clusterings and include all the other dissimilar ones, for avoid catastrophic forgetting. In the final clustering, one can distinguish, among the groups, anomalous behaviors of the turbines and relate them to specific root causes, such that for a new coming *NPP* turbine, one can estimate and predict its future evolution, reducing the time for fault detection and diagnosis and facilitating the decision on how to intervene to avoid the consequences of the fault.

The proposed framework has been applied to 149 and 116 shut-down transients of two *NPPs* turbines, respectively. Three alternative approaches for learning data have also been applied and their results have been compared to those obtained by the proposed framework.

The comparison of the results obtained show that the proposed framework: 1) is capable of incrementally learning the behaviors of the turbines fleet under varying environmental and operational conditions, 2) requires less computational efforts by avoiding to retrain each time on all shut-down transients that have been collected thus far from the turbines fleet. The performance of the final obtained consensus clusters, quantified in terms of clusters separation and compactness, is satisfactory.

Further improvement in the clustering procedure and automation of the identification of the best matching cluster will be object of future work.

Acknowledgements

The participation of Sameer Al-Dahidi and Piero Baraldi to this research is supported by the European Union Project INNovation through Human Factors in risk analysis and management (INNHF, www.innhf.eu) funded by the 7th framework program FP7-PEOPLE-2011-Initial Training Network: Marie-Curie Action. The participation of Enrico Zio to this research is partially supported by the China NSFC under grant number 71231001.

The authors would like to thank all the reviewers for their valuable comments to improve the quality of this paper.

Appendix A. The unsupervised ensemble clustering

The unsupervised ensemble clustering approach is proposed to 1) handle the missing data in the original dataset, and 2) avoid the need of having an "a priori" knowledge of the number of clusters M in the final consensus clustering.

The flowchart for the method is sketched in Fig. A1. The method goes along the following steps:

Step 1: *Adjacency matrix computation*. An adjacency binary similarity matrix $\bar{\bar{A}}$, is built by aggregating the similarities μ of the H base clusterings [60], where for each j-th base clustering, the similarity $\mu = 1$, if two data belong to the same cluster, whereas the similarity $\mu = 0$, if they belong to different clusters.

Step 2: *Similarity matrix computation*. From the adjacency binary similarity matrix $\bar{\bar{A}}$, the overall similarity matrix $\bar{\bar{S}}$ is computed as the entry-wise average of the H base clusterings, i.e. $\bar{\bar{S}} = \left(1/H\right)\bar{\bar{A}}\bar{\bar{A}}^T$ [60]. In this way, each entry of the similarity matrix has a value in [0, 1], which is proportional to how likely a pair of data is, when grouped together.

Step 3: *Spectral Clustering*. Once the overall similarity matrix $\bar{\bar{S}}$ is computed, Spectral Clustering (Appendix B) is used to reveal the hidden structure of $\bar{\bar{S}}$. The basic idea of Spectral Clustering is to extract the relevant information of the matrix $\bar{\bar{S}}$, by considering the eigenvectors associated to the ascended eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{C_{Candidate}}, \ldots, \lambda_N$ of the normalized laplacian matrix $\bar{\bar{L}}_{rs}$ of $\bar{\bar{S}}$, to perform dimensionality reduction before clustering in fewer dimensions (see Step 1 in Appendix B) [30,61]. The eigenvectors $\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_{C_{Candidate}}, \ldots, \bar{u}_N$ of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{C_{Candidate}}, \ldots, \lambda_N$ are calculated and stored in a matrix $\bar{\bar{U}}$ with a size NxN (see Steps 2 and 4 in Appendix B), where $C_{Candidate} = [C_{min}, C_{max}]$ and C_{min} and C_{max} are the minimum and maximum numbers of clusters considered for the final consensus clustering, respectively.

Step 4: *Clustering algorithm*. For each candidate number of clusters $C_{Candidate}$, the reduced matrix of $\bar{\bar{U}}$ with a size Nx$C_{Candidate}$ is partitioned into $C_{Candidate}$ clusters by using a single clustering algorithm and the final consensus clustering $P^*_{C_{Candidate}}$ is obtained. In this work, we resort to the K-means algorithm, one of the most used clustering methods, to partition $\bar{\bar{C}}$ into $K = C_{Candidate}$ clusters [62,63].

Step 5: *Final consensus clustering selection*. For each $\boldsymbol{C_{Candidate}}$, the obtained consensus clustering $\boldsymbol{P^*_{C_{Candidate}}}$ is evaluated by computing its Silhouette validity index $\boldsymbol{SV_{C_{Candidate}}}$ [34]. The most appropriate consensus clustering $\boldsymbol{P^*_{C^*}}$ is the one for which the Silhouette reaches a maximum, for which clusters are well separated and compacted (see also Appendix C).

Appendix B. Unsupervised spectral clustering

Spectral clustering technique uses the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions [30,61]. In this work, the similarity matrix $\bar{\bar{S}}$ of size NxN is computed by Cluster-based Similarity Partitioning Algorithm (*CSPA*). The Spectral Clustering technique entails four steps [30,61]:

Step 1: *Normalized Laplacian Matrix*. Starting from the similarity matrix $\bar{\bar{S}}$, the degree matrix $\bar{\bar{D}}$ is calculated, whose entries $d_1, d_2, \ldots, d_N$ are:

$$d_i = \sum_{j=1}^{N} S_{ij}, i = 1, 2, \ldots, N \tag{A1}$$

Based on $\bar{\bar{D}}$, the normalized Laplacian matrix $\bar{\bar{L}}_{rs}$, is calculated:

$$\bar{\bar{L}}_{rs} = \bar{\bar{D}}^{-1}\bar{\bar{L}} = \bar{\bar{I}} - \bar{\bar{D}}^{-1}\bar{\bar{S}} \tag{A2}$$

where $\bar{\bar{L}} = \bar{\bar{D}}\text{-}\bar{\bar{S}}$ and $\bar{\bar{I}}$ is the identity matrix of size $[N, N]$.

Step 2: *Eigenvalues and eigenvectors of* $\bar{\bar{L}}_{rs}$. Given $\bar{\bar{L}}_{rs}$, compute the eigenvectors $\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_N$. The first C eigenvalues are such that they are very small whereas λ_{C+1} is relatively large [64].

Step 3: *Number of clusters*. The number of clusters is set equal to C, according to the eigengap heuristic theory [64].

Step 4: *Feature extraction*. The relevant information on the structure of the matrix $\bar{\bar{S}}$ is obtained by considering the eigenvectors $\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_N$ associated to the C smallest eigenvalues of its laplacian matrix $\bar{\bar{L}}_{rs}$. The square matrix $\bar{\bar{S}}$ is transformed into a matrix

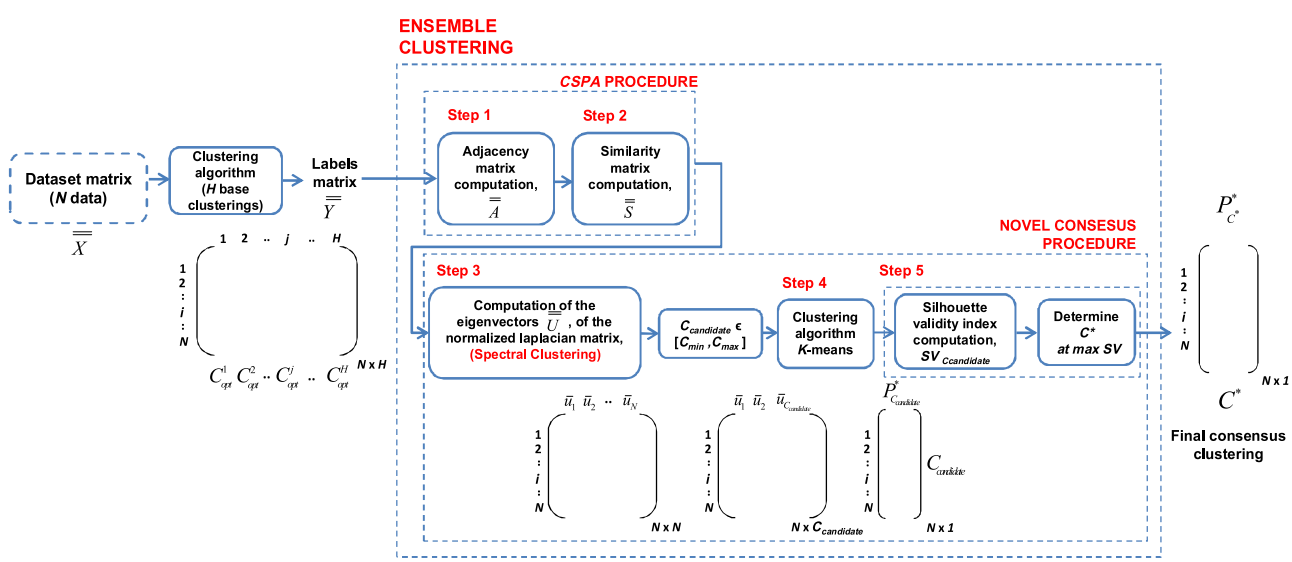

Fig. A1. Flowchart of the ensemble clustering approach.

$\overline{\overline{U}}$ of size $[N, C]$, in which the C columns of $\overline{\overline{U}}$ are the eigenvectors [61].

Appendix C. Silhouette validity index

To evaluate the optimal number of clusters C^* among several clusters candidates, Silhouette validity index has been adopted. The silhouette value for the i-th datum, $i = 1, \ldots, N$, is a measure of how similar/dissimilar that datum is to others in its own cluster and to the other clusters, respectively. The silhouette value for the i-th datum S^i is defined as [34]:

$$S^i = (b_i - a_i)/\max(a_i, b_i) \tag{A3}$$

where a_i is the average distance from the i-th datum to the others in the same cluster, and b_i is the minimum average distance from the i-th datum to the others in a different cluster, minimized over clusters.

The mean of the silhouette values for the m-th cluster C_m is called the cluster mean silhouette and is denoted as S_m (Eq. (A4)):

$$S_m = \frac{1}{n_m} \sum_{i \in C_m} S^i \tag{A4}$$

where n_m is total number of data in the m-th cluster. Finally, the global silhouette index $SV_{C_{Candidate}}$ is the mean of the mean silhouettes (Eq. (A5)) through all the clusters.

$$SV_{C_{Candidate}} = \frac{1}{C_{Candidate}} \sum_{m=1}^{C_{Candidate}} S_m \tag{A5}$$

The silhouette value ranges from −1 to +1. A high silhouette value SV_{C^*} indicates that the C^* clusters of the final consensus clustering are well separated and compacted.

Appendix D. Sensitivity analysis of the *K* value on the quality of the updated consensus clusters

Building the *FKNN* classifier for allocating the new coming EE1 transients to the existing FF1 consensus clusters requires to optimally set the K nearest neighbors value. In fact, neither a too small nor a too large value of K can be considered as a valuable result from the practical point of view of assigning new transients of EE1 to the available FF1 consensus clusters: a small value of K leads to over fitting the data and accordingly to higher variance in the classification task (i.e., classifier is less stable), whereas a large value of K leads to under fitting the data and accordingly to higher bias in the classification task (i.e., classifier is less precise) [65]. Therefore, an optimum value of K needs to be identified.

In this analysis, the optimum K value, K^*, is selected among different values of K, $K_{Candidate}$, that span in the interval $[K_{min}, K_{max}]$, where K_{min} is the minimum number of nearest neighbors that is usually set to 2 and K_{max} is the maximum number of nearest neighbors that is usually set to the square root of the size of the training set $\overline{\overline{X}}_{FF1}$ (i.e., 13), by evaluating quality of the updated FF1 consensus clusters, in terms of clusters separation and compactness: the optimum K value is the value at which the Silhouette measure is maximized, while both the *C-index* and *DB* measures are minimized, which makes the selection process a multi-criteria decision problem.

To solve this, we resort to the Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) method [66,67], which is a multiple criteria decision making method whose basic principle is that the selected solution should have the shortest distance from the ideal solution and the farthest distance from the negative ideal solution. In principle, different criteria can have different weights (i.e., importance) when selecting the solution, depending on the objective of the study (refer to [66,67], for more details). In this study, since there is no evidence on the relative importance of the three selected criteria, i.e., Silhouette measure, *C-index* and *DB* value, they have been assigned equal weights, leading to an optimum K value $K^* = 7$ for which the Silhouette measure is −0.0230, *C-index* is 0.5735 and *DB* value is 3.2235.

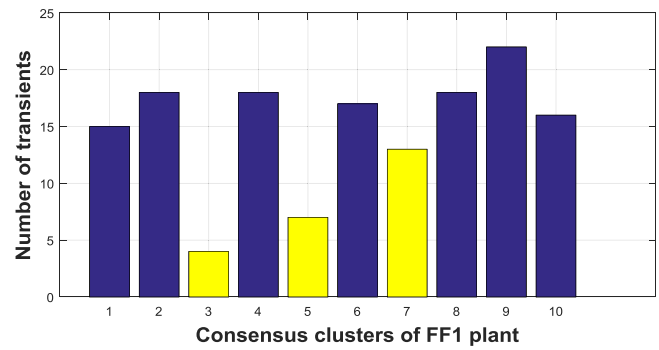

Fig. E1. The distributive characteristics of the available FF1 consensus clusters.

Appendix E. Investigation of the distributive characteristics of FF1 consensus clusters on the quality of the updated clusters

Another issue needs to be consider for building the *FKNN* classifier is the distributive characteristics of the available training dataset $\overline{\overline{X}}_{FF1}$. In fact, most standard algorithms used for classification assume balanced class distributions, i.e., equal proportions of number of instances in the available classes. However, in real-world applications, dataset is usually imbalanced, that is the number of instances in one class (majority class) is much larger than that in another class (minority class), that makes the algorithms biased towards the majority classes and therefore there is a higher misclassification rate for the minority class instances [59].

In the case study under analysis, Fig. E1 shows the number of transients in the $P^*_{FF1} = 10$ consensus clusters. One can consider that seven clusters (dark shade of color) comprises most of the transients (i.e., 124 transients) compare to the remaining clusters (i.e., 24 transients) (light shade of color). The imbalance ratio is ∼5, i.e., the dataset is low imbalanced.

To tackle this issue, the ADAptive SYNthetic (ADASYN) sampling approach [68] has been adopted for learning from the imbalanced dataset $\overline{\overline{X}}_{FF1}$. The basic idea of ADASYN is to balance the data sizes in majority and minority classes by generating more synthetic data for minority class instances in the vicinity of the boundary between the two classes (refer to [68] for more details). The updated training dataset $\overline{\overline{X}}^*_{FF1}$ will be used for training the *FKNN* classifier.

In this regard, the sensitivity analysis of Appendix D for the selection of the optimum K value is then repeated considering the updated training dataset $\overline{\overline{X}}^*_{FF1}$. For each K candidate, the quality of the updated FF1 consensus clusters is calculated, in terms of clusters separation and compactness. The optimum K values is found to be at $K^* = 7$ by resorting to the TPOSIS method for which the Silhouette measure is 0.0659, *C-index* is 0.4687 and *DB* measure is 3.0367.

One can notice that the optimum K value is still the same as already found when the original training dataset $\overline{\overline{X}}_{FF1}$ is used, and the goodness of the final clusters is slightly enhanced, but still comparable to the quality of the final clusters found by the proposed approach (Table 7 and Fig. 11). This can be justified by the fact that the imbalanced ratio of the original training dataset is low (i.e., 1:5).

References

[1] S. Katipamula, M. Brambley, Review article methods for fault detection, diagnostics, and prognostics for building systems—a review, part I, HVAC&R Res. 11 (2005) 3–25, http://dx.doi.org/10.1080/10789669.2005.10391123.

[2] V. Agarwal, N.J. Lybeck, R. Bickford, R. Rusaw, Development of asset fault signatures for Prognostic and Health Management in the nuclear industry, 2014 Int. Conf. Progn. Heal. Manag., IEEE (2014) 1–7, http://dx.doi.org/10.1109/ICPHM.2014.7036366.

[3] G. Medina-Oliva, A. Voisin, M. Monnin, J.-B. Leger, Predictive diagnosis based on a fleet-wide ontology approach, Knowl.-Based Syst. 68 (2014) 40–57, http://dx.doi.org/10.1016/j.knosys.2013.12.020.

[4] A.K.S. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance, Mech. Syst. Signal Process. 20 (2006) 1483–1510, http://dx.doi.org/10.1016/j.ymssp.2005.09.012.

[5] R. Patrick, M.J. Smith, C.S. Byington, G.J. Vachtsevanos, K. Tom, C. Ly, Integrated software platform for fleet data analysis, enhanced diagnostics, and safe transition to prognostics for helicopter component CBM, in: Proceedings of Annual Conference of the Prognostics and Health Management Society, Portland, Oregon, October 10-16, 2010, pp. 1–15.

[6] Y. Lei, Z. He, Y. Zi, A new approach to intelligent fault diagnosis of rotating machinery, Expert Syst. Appl. 35 (2008) 1593–1600, http://dx.doi.org/10.1016/j.eswa.2007.08.072.

[7] L. Batista, B. Badri, R. Sabourin, M. Thomas, A classifier fusion system for bearing fault diagnosis, Expert Syst. Appl. 40 (2013) 6788–6797, http://dx.doi.org/10.1016/j.eswa.2013.06.033.

[8] V. Agarwal, N.J. Lybeck, B.T. Pham, R. Rusaw, R. Bickford, Online monitoring of plant assets in the nuclear industry, Annu. Conf. PHM Soc. (2013) 1–12.

[9] Z. Su, B. Tang, L. Deng, Z. Liu, Fault diagnosis method using supervised extended local tangent space alignment for dimension reduction, Measurement 62 (2015) 1–14, http://dx.doi.org/10.1016/j.measurement.2014.11.003.

[10] International Atomic Energy Agency (IAEA), On line Monitoring for improving Performance of Nuclear power plants, in: Part 2: Process and component condition monitoring and diagnostics, IAEA Nucl. Energy Ser., 2013.

[11] J. Ma, J. Jiang, Applications of fault detection and diagnosis methods in nuclear power plants: a review, Prog. Nucl. Energy 55 (2011) 255–266.

[12] R. Isermann, Process fault detection based on modeling and estimation methods-a survey, Automatica 20 (1984) 387–404, http://dx.doi.org/10.1016/0005-1098(84)90098-0.

[13] V. Venkatasubramanian, R. Rengaswamy, K. Yin, S.N. Kavuri, A review of process fault detection and diagnosis part I: quantitative model-based methods, Comput. Chem. Eng. 27 (2003) 293–311, http://dx.doi.org/10.1016/S0098-1354(02)00160-6.

[14] Y. Ben-Haim, Malfunction isolation in linear stochastic systems: application to nuclear power plants, Nucl. Sci. Eng. 85 (1983) 156–166.

[15] E. Turkcan, O. Ciftcioglu, T.H.J.J. van der Hagen, Surveillance and fault diagnosis for power plants in the Netherlands: operational experience, in: Proc IAEA Tech. Comm. Meet. Diagnostic Syst. Nucl. Power Plants, Istanbul, Turkey, 1998, pp. 53–70.

[16] R.J. Patton, J. Chen, Observer-based fault detection and isolation: robustness and applications, Control Eng. Pract. 5 (1997), http://dx.doi.org/10.1016/S0967-0661(97)00049-X.

[17] J. Reifman, Survey of artificial intelligence methods for detection and identification of component faults in nuclear power plants, Nucl. Technol. 119 (1997) 76–97.

[18] H.M. Hashemian, On-line monitoring applications in nuclear power plants, Prog. Nucl. Energy 53 (2011) 167–181, http://dx.doi.org/10.1016/j.pnucene.2010.08.003.

[19] V. Venkatasubramanian, A review of process fault detection and diagnosis: part III: process history based methods, Comput. Chem. Eng. 27 (2003) 293–311, http://dx.doi.org/10.1016/S0098-1354(02)00162-X.

[20] G. Zhou, L. Yang, Advance in study of intelligent diagnostic method for nuclear power plant, Yuanzineng Kexue Jishu/Atomic Energy Sci. Technol. 42 (2008) 92–99.

[21] M. Todd, S.D.J. McArthur, J.R. McDonald, S.J. Shaw, A semiautomatic approach to deriving turbine generator diagnostic knowledge, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 37 (2007) 979–992, http://dx.doi.org/10.1109/TSMCC.2007.900645.

[22] M.J. Embrechts, S. Benedek, Hybrid identification of nuclear power plant transients with artificial neural networks, IEEE Trans. Ind. Electron. 51 (2004) 1438–1443, http://dx.doi.org/10.1109/TIE.2004.824874.

[23] M. Monnin, B. Abichou, A. Voisin, C. Mozzati, Fleet historical cases for predictive maintenance, in: Int Conf. Surveill. 6, Compiègne 6, 2010, pp. 25–26.

[24] T. Wang, J. Yu, D. Siegel, J. Lee, A similarity-based prognostics approach for Remaining Useful Life estimation of engineered systems, PHM. Int. Conf. (2008) 1–6.

[25] E. Olsson, P. Funk, M. Bengtsson, Fault Diagnosis of Industrial Robots Using Acoustic Signals and Case-Based Reasoning, 2004, pp. 686–701.

[26] A. Reymonet, J. Thomas, N. Aussenac-Gilles, Ontology based information retrieval: an application to automotive diagnosis, Proc. Int. Work Princ. Diagn. (2009) 9–14.

[27] P. Umiliacchi, D. Lane, F. Romano, Predictive Maintenance of railway subsystems using an Ontology based modelling approach, in: Proc. 9th World Conf. Railw. Res., Lille, France, 2011, pp. 1–10.

[28] L. Serir, E. Ramasso, N. Zerhouni, Evidential evolving Gustafson-Kessel algorithm for online data streams partitioning using belief function theory, Int. J. Approx. Reason. 53 (2012) 747–768, http://dx.doi.org/10.1016/j.ijar.2012.01.009.

[29] L. Serir, E. Ramasso, P. Nectoux, N. Zerhouni, E2GKpro. An evidential evolving multi-modeling approach for system behavior prediction with applications, Mech. Syst. Signal Process. 37 (2013) 213–228, http://dx.doi.org/10.1016/j.ymssp.2012.06.023.

[30] P. Baraldi, F. Di Maio, M. Rigamonti, E. Zio, R. Seraoui, Unsupervised clustering of vibration signals for identifying anomalous conditions in a nuclear turbine, J. Intell. Fuzzy Syst. 28 (2013) 1723–1731, http://dx.doi.org/10.3233/IFS-141459.

[31] F. Wu, J. Lee, Information reconstruction method for improved clustering and diagnosis of generic gearbox signals, Int. J. Progn. Heal. Manage. 2 (2011) 1–9.

[32] S. Al-Dahidi, F. Di Maio, P. Baraldi, E. Zio, R. Seraoui, A novel ensemble clustering for operational transients classification with application to a nuclear power plant turbine, Int. J. Progn. Heal. Manage. 6 (2015) 1–21.

[33] S. Al-Dahidi, F. Di Maio, P. Baraldi, E. Zio, R. Seraoui, Unsupervised ensemble clustering for transients classification in a nuclear power plant turbine, Saf Reliab. Methodol. Appl. – Proc. ESREL Conf. (2015) 2339–2347.

[34] P.J. Rousseeuw, Silhouettes, A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, http://dx.doi.org/10.1016/0377-0427(87)90125-7.

[35] L. Hubert, J. Schultz, Quadratic assignment as a general data analysis strategy, Br. J. Math. Stat. Psychol. 29 (1976) 190–241, http://dx.doi.org/10.1111/j.2044-8317.1976.tb00714.x.

[36] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1979) 224–227, http://dx.doi.org/10.1109/TPAMI.1979.4766909.

[37] E. Zio, F. di Maio, Fault diagnosis and failure mode estimation by a data-driven fuzzy similarity approach, Int. J. Perform. Eng. 8 (2012) 49–65.

[38] A. Joentgen, L. Mikenina, R. Weber, H.-J. Zimmermann, Dynamic fuzzy data analysis based on similarity between functions, Fuzzy Sets Syst. 105 (1999) 81–90, http://dx.doi.org/10.1016/S0165-0114(98)00337-6.

[39] M.G. Na, S.H. Shin, S.M. Lee, D.W. Jung, S.P. Kim, J.H. Jeong, B.C. Lee, Prediction of major transient scenarios for severe accidents of nuclear power plants, IEEE Trans. Nucl. Sci. 51 (2004) 313–321, http://dx.doi.org/10.1109/TNS.2004.825090.

[40] R. Östermark, A fuzzy vector valued KNN-algorithm for automatic outlier detection, Appl. Soft Comput. J. 9 (2009) 1263–1272, http://dx.doi.org/10.1016/j.asoc.2009.03.009.

[41] H.L. Chen, C.C. Huang, X.G. Yu, X. Xu, X. Sun, G. Wang, S.J. Wang, An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach, Expert Syst. Appl. 40 (2013) 263–271, http://dx.doi.org/10.1016/j.eswa.2012.07.014.

[42] J.M. Keller, M.R. Gray, A fuzzy K-nearest neighbor algorithm, IEEE Trans. Syst. Man Cybern. SMC-15 (1985) 580–585, http://dx.doi.org/10.1109/TSMC.1985.6313426.

[43] E. Rendón, I. Abundez, A. Arizmendi, E.M. Quiroz, Internal versus External cluster validation indexes, Int. J. Comput. Commun. 5 (2011) 27–34.

[44] International Atomic Energy Agency (IAEA), Advanced Surveillance, Diagnostics, and Prognostics Techniques Used for Health Monitoring of Systems, Structures, and Components in NPPs, Nucl. Energy Ser, 2011.

[45] K. Salahshoor, M. Kordestani, M.S. Khoshro, Fault detection and diagnosis of an industrial steam turbine using fusion of SVM (support vector machine) and ANFIS (adaptive neuro-fuzzy inference system) classifiers, Energy 35 (2010) 5472–5482, http://dx.doi.org/10.1016/j.energy.2010.06.001.

[46] E. Zio Prognostics and, Health management of industrial equipment, Diagn. Progn. Eng. Syst. Methods Tech. IGI-Global 2012 (2012) 333–356, http://dx.doi.org/10.4018/978-1-4666-2095-7.ch017.

[47] M.E. Sharp, J.W. Hines, R. Austin, Equipment monitoring via transient methods EPRI technology innovation program, in: 7th Int. Top. Meet. Nucl. Plant Instrumentation, Control. Human-Machine Interface Technol. 2010, NPIC HMIT 2010, 2010, pp. 1207–1219.

[48] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer-Verlag, New York, 2005, http://dx.doi.org/10.1007/b98888.

[49] J. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, New York, 2002.

[50] R. Polikar, Ensemble based systems in decision making, Circuits Syst. Mag. IEEE 6 (2006) 21–45.

[51] C. Liu, G.F. Wang, Z.M. Li, Incremental learning for online tool condition monitoring using Ellipsoid ARTMAP network model, Appl. Soft Comput. 35 (2015) 186–198, http://dx.doi.org/10.1016/j.asoc.2015.06.023.

[52] C.M. Rocco, S.E. Zio, A support vector machine integrated system for the classification of operation anomalies in nuclear components and systems, Reliab. Eng. Syst. Saf. 92 (2007) 593–600, http://dx.doi.org/10.1016/j.ress.2006.02.003.

[53] A. Ng, M.I. Jordan, On generative vs. discriminative classifiers: a comparison of logistic regression and naive bayes, Proc. Adv. Neural Inf. Process. 28 (2002) 169–187, http://dx.doi.org/10.1007/s11063-008-9088-7.

[54] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Taylor & Francis, 1984.

[55] Y. Guo, T. Hastie, R. Tibshirani, Regularized Linear Discriminant Analysis and Its Application in Microarrays, 2007, http://dx.doi.org/10.1093/biostatistics/kxj035.
[56] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction, Math. Intell. 27 (2005) 83–85, http://dx.doi.org/10.1007/BF02985802.
[57] F. Di Maio, P. Secchi, S. Vantini, E. Zio, Fuzzy C-means clustering of signal functional principal components for post-processing dynamic scenarios of a nuclear power plant digital instrumentation and control system, IEEE Trans. Reliab. 60 (2011) 415–425, http://dx.doi.org/10.1109/TR.2011.2134230.
[58] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. 36 (1974) 111–147, http://dx.doi.org/10.2307/2984809.
[59] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. (NY). 250 (2013) 113–141, http://dx.doi.org/10.1016/j.ins.2013.07.007.
[60] A. Strehl, J. Ghosh, Cluster ensembles – a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2002) 583–617, http://dx.doi.org/10.1162/153244303321897735.
[61] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (2007) 395–416, http://dx.doi.org/10.1007/s11222-007-9033-z.
[62] M. Su, C. Chou, A modified version of the K-means algorithm with a distance based on cluster symmetry, Pattern Anal. Mach. Intell. 23 (2001) 674–680, http://dx.doi.org/10.1109/34.927466.
[63] X.Z. Fern, W. Lin, Cluster ensemble selection, Stat. Anal. Data Min. 1 (2008) 787–797, http://dx.doi.org/10.1002/sam.10008.
[64] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, Nips 14 (2001) 849–856.
[65] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21–27, http://dx.doi.org/10.1109/TIT.1967.1053964.
[66] S. Opricovic, G.H. Tzeng, Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS, Eur. J. Oper. Res. 156 (2004) 445–455, http://dx.doi.org/10.1016/S0377-2217(03)00020-1.
[67] C.-L. Hwang, K. Yoon, Methods for Multiple Attribute Decision Making, Springer, Berlin Heidelberg, 1981, http://dx.doi.org/10.1007/978-3-642-48318-9_3.
[68] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263–1284, http://dx.doi.org/10.1109/TKDE.2008.239.