Genetic Algorithm-based Wrapper Approach for Grouping Condition Monitoring Signal of Nuclear Power Plant Components

Piero Baraldi, Roberto Canesi, Enrico Zio, Rédouane Seraoui, Roger Chevalier

Genetic algorithm-based wrapper approach for grouping condition monitoring signals of nuclear power plant components

Piero Baraldi[a,*], Roberto Canesi[a], Enrico Zio[b,a], Redouane Seraoui[c] and Roger Chevalier[c]
[a]*Dipartimento di Energia, Politecnico di Milano, viale Ponzio 34/3, Milano, Italy*
[b]*Ecole Centrale Paris – Supelec, Paris, France*
[c]*EDF - R&D \ STEP Simulation et Traitement de l'information pour l'exploitation des systèmes de production, 6 quai Watier, Chatou Cedex, France*

Abstract. Equipment condition monitoring of nuclear power plants requires to optimally group the usually very large number of signals and to develop for each identified group a separate condition monitoring model. In this paper we propose an approach to optimally group the signals. We use a Genetic Algorithm (GA) for the optimization of the groups; the decision variables of the optimization problem relate to the composition of the groups (i.e., which signals they contain) and the objective function (fitness) driving the search for the optimal grouping is constructed in terms of quantitative indicators of the performances of the condition monitoring models themselves: in this sense, the GA search engine is a wrapper around the condition monitoring models. A real case study is considered, concerning the condition monitoring of the Reactor Coolant Pump (RCP) of a Pressurized Water Reactor (PWR). The optimization results are evaluated with respect to the accuracy and robustness of the monitored signals estimates. The condition monitoring models built on the groups found by the proposed approach outperform the model which uses all available signals, whereas they perform similarly to the models built on groups based on signal correlation. However, these latter do not guarantee the robustness of the reconstruction in case of abnormal conditions and require to a priori fix characteristics of the groups, such as the desired minimum correlation value in a group.

Keywords: Condition Monitoring, Abnormal Condition Detection, Genetic Algorithm search, Reactor Coolant Pump, Pressurized Water Reactor

[*]Corresponding author. E-mail: piero.baraldi@polimi.it, Tel. +39 02 2399 6355, Fax. +39 02 2399 6309

1. Introduction

Monitoring the condition of a component can be based on a (typically empirical) model estimating the values of measurable variables (signals) in normal conditions. During operation, the measured values are compared with those estimated ('reconstructed') by the model: a deviation between the observed and reconstructed values reveals the presence of an abnormal condition [30].

In practical industrial implementations, the performance of a single model estimating all the signals measured by the sensors, usually a very large number, may not be satisfactory [31]. In [3] it has been shown that grouping the signals and then building a specialized model for each group allows to remarkably increase the condition monitoring performance. With respect to the structure of the groups, different approaches have been proposed. In [7] a procedure which finds the optimal group for the reconstruction of a single target signal has been considered. According to this approach, the grouping problem has been reduced to the problem of selecting those signals which allow to obtain the best reconstruction of the target signal. Notice, however, that if one is interested in the reconstruction of all the available signals, the search of the optimal group has to be repeated for all the signals and thus the grouping results very demanding from the computational point of view.

Alternatively, grouping techniques based on overlapping groups (i.e. the same signal can belong to more than one group) have been proposed in [8-11]. In these approaches, the number of groups and the number of signals in each group are a priori fixed according to the user needs and a feature subset selection algorithm is developed to provide the optimal grouping. Then, a specialized reconstruction model for each group is built, and the obtained ensemble of models can be used for the signal reconstruction. Since the same signal belongs to several groups, the outcomes of the individual models containing the signal are properly combined to produce the final reconstruction.

Non overlapping, mutually exclusive groups have been considered in [3,35]. In these works, all signals are assigned to exactly one group. In this way, the number of models to be developed and the computational efforts in the signal reconstruction phase are reduced with respect to the ensemble approaches. For this reason, in real industrial applications a grouping structure based on non overlapping group tends to be preferred.

With respect to the criteria which can be used to decide which signals should be assigned to each non overlapping group, filter and wrapper approaches can be followed.

In filter approaches signal grouping is based on characteristics judged to be (indirectly) favorable for condition monitoring. In this respect, several criteria for signal grouping have been investigated, e.g., the location of the measurements (i.e., signals measured in the same area of the plant are put in the same group), their correlation (i.e., the groups are formed by correlated signals), and others. Tests on a real case study have shown the superior performance of a correlation-based grouping between these criteria [6]. Notice that once the criterion for the grouping has been fixed (e.g. correlation), it is necessary to find the optimal groups with respect to that criterion. To this purpose, the authors have proposed an heuristic grouping approach which requires to a priori fix a correlation threshold: signals with an absolute value of the correlation coefficient larger than the threshold are put in the same group [6]. Similarly, a filter method for the identification of optimal groups for multivariate time series analysis has been proposed in [35]. The method is based on the definition of a partition metric measuring how much the groups are formed by signals highly correlated with signals of the same group and lowly correlated with signals of other groups. Then, various methods such as Genetic Algorithms (GAS) and Hill Climbing (HC) have been compared in order to find out the grouping which maximizes the partition metric. The performance of the filter approaches in [6,35] has been shown to be good in a real case study, although the groups are selected based on the correlation which is an indirect signal characteristic, independent from the reconstruction algorithm actually used.

In wrapper approaches, a search algorithm is used as a "wrapper" around the condition monitoring model (Figure 1); during the optimization search, the performance of the condition monitoring model itself is directly used as an evaluation function to compare the different groups selected by the search engine 25, 36]. Wrapper approaches are expected to be more performing than the filter ones since in the former the groups of signals found are optimal for the specific reconstruction model used, i.e., different condition monitoring models, such as Auto-Associative Kernel Regression (AAKR) or Principal Component Analysis (PCA), would, in general, lead to different optimal groups [18]. On the other side, wrapper approaches are generally computationally less efficient than the filter approaches because for each grouping of trial,

the development of a complete reconstruction model computation is more time consuming than the computation of an evaluation function from the available data.

In this work we investigate the possibility of using a wrapper approach for the search of the optimal signal groups, considering GAs as search engines. The motivation of the choice of the GAs is found in their ability of finding the optimal solution by efficiently scanning the search space in an acceptable computational time. In particular, GAs have been applied with success to many different optimization problems such as pattern recognition [26,36], machine learning [21], maintenance planning [14,23], lifecycle cost optimization [32], production optimization [4], highway alignment optimization [22], robotics [33], the optimization of traffic control signals [34], electrical transmission towers [28] and large structures [1-3,23].

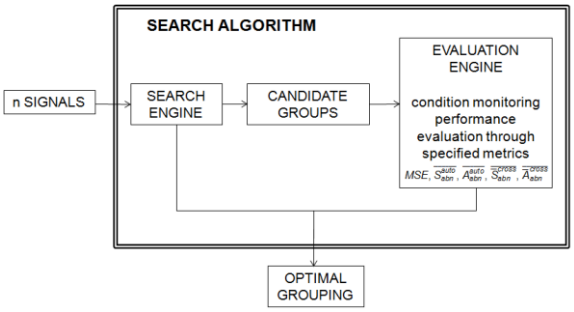

Fig. 1. Wrapper approach for optimal signal grouping.

In the present work, the GA-based wrapper approach is tested on a real case study concerning 46 signals selected between those used to monitor the reactor coolant pump of a Pressurized Water Reactor (PWR). The condition monitoring performance is evaluated with respect to metrics that measure i) the accuracy, i.e., the ability of the overall model to correctly and accurately reconstruct the signal values when the plant is in normal operation; ii) the robustness, i.e., the overall model ability to reconstruct the signal values in case of abnormal operation and consequent anomalous behavior of some monitored signals [19]. The results are compared with those achieved by considering a single group formed by all signals and by groups based on signal correlation.

2. Condition Monitoring

Figure 2 shows a typical scheme of condition monitoring of a component. Sensor measurements $\vec{x}^{obs}$ are sent to an auto-associative empirical model of the component behavior in normal condition (nc). Thus, the model provides in output the values expected in case of normal condition, $\vec{\hat{x}}_{nc}$, of the input signals. A deviation between the measured $\vec{x}^{obs}$ and reconstructed $\vec{\hat{x}}_{nc}$ values in one or more signals reveals the presence of faults [30].

In other words, in case of normal condition, the measured value $\vec{x}^{obs}$ should be very similar to the model reconstructions $\vec{\hat{x}}_{nc}$, whereas in case of abnormal condition (ac) the model still reconstructs $\vec{\hat{x}}_{nc}$, which differs from the measured values $\vec{x}^{obs}$. Notice that one usually does not know whether the component is working in normal or abnormal conditions, whereas, by observing the residuals $\vec{r} = \vec{x}^{obs} - \vec{\hat{x}}_{nc}$, it is possible to detect the component condition. In this respect, several methods of analysis of the residuals $\vec{r}$ for fault detection exist, e.g. the Sequential Probability Ratio Test (SPRT) [8,18].

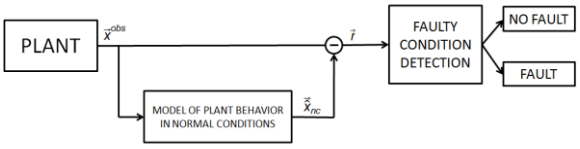

Fig. 2. Condition monitoring scheme.

2.1. Auto-Associative Kernel Regression (AAKR) method

The model considered in this work for reconstructing the component behavior in normal condition is based on the AAKR method [13]. AAKR is an empirical modeling technique that uses historical observations of the signals taken during normal plant operation. The basic idea of the method is to reconstruct the signal values in case of normal condition, $\vec{\hat{x}}_{nc}$, given a current signal measurement vector, $\vec{x}^{obs} = (x^{obs}(1), \ldots, x^{obs}(n))$, as a weighted sum of the historical observations. Appendix A provides the details of the method.

If the performance of a single model estimating all signals measured by the sensors is not satisfactory [31], several AAKR reconstruction models can be built, each one estimating only a subset of the signals in $\vec{x}^{obs}$.

3. Condition monitoring performance metrics

In order to evaluate the performance of a condition monitoring model, the following criteria should be considered.

1) The accuracy, i.e. the ability of the model to correctly and accurately reconstruct the signal values when the plant is in normal operation. An accurate condition monitoring model allows to reduce the number of false alarms, i.e. detections of faulty behaviors when no faulty conditions are actually occurring.

The accuracy metric is typically defined as the mean square error between the model reconstruction and the signal measured values. Let $X^{obs\text{-}nc}$ be a matrix of observed data whose generic element $X^{obs\text{-}nc}(k,j)$ represents the k-th time observation, k=1,…,N, of the j-th measured signal, j=1,…,n, taken during normal plant condition, $X^{\text{test-nc}}$ a matrix of signal measurements different from those in $X^{\text{obs-nc}}$, with $X^{\text{test-nc}}(k,j)$ indicating the true value of the j-th signal, j=1,…,n, of the k-th test pattern, k=1,…,N^{test} and $\hat{X}_{\text{nc}}^{\text{test}}(k,j)$ its reconstruction provided by the condition monitoring model; then, the mean square error with respect to signal j is [19]:

$$MSE(j) = \frac{\sum_{k=1}^{N^{test}} (\hat{X}_{nc}^{test-nc}(k,j) - X_{nc}^{test-nc}(k,j))^2}{N^{test}} \quad (1)$$

A global accuracy measure that takes into account all the monitored signals and test patterns is defined by:

$$MSE = \frac{\sum_{j=1}^{n}\sum_{k=1}^{N^{test}} \frac{(\hat{X}_{nc}^{test-nc}(k,j) - X_{nc}^{test-nc}(k,j))^2}{N^{test}}}{n} = \frac{\sum_{j=1}^{n} MSE(j)}{n} \quad (2)$$

Notice that, although the metric is named accuracy, it is actually a measure of error and, thus, a low value is desired.

2) The robustness, i.e. the model ability to reconstruct the values of the signals of interest in abnormal operation when some monitored signals behave anomalously. In abnormal plant conditions, a robust model reconstructs the value of a plant signal as if the plant were in normal operation: then, the differences between the measured and the reconstructed signal values can easily identify the abnormal condition.

In this respect, real data measured by the sensors in abnormal plant conditions are usually not available because these latter are rare; then simulation is used to artificially inject abnormality by adding realistic deviations to the signals measured during normal plant operation. Let $X^{\text{test-ac}(i)}$ be a matrix of test patterns whose values of the i-th signal have been disturbed with deviations, with $X^{\text{test-ac}(i)}(k,j)$ indicating the value of the j-th signal of the k-th test pattern, k=1,…,N^{test}, and $\hat{X}_{\text{nc}}^{\text{test-ac}(i)}(k,j)$ its reconstruction provided by the condition monitoring model which is expected to be the signal value in normal condition $X^{\text{test-nc}}(k,j)$.

Quantitative indicators of robustness can then be introduced as follows.

The auto-sensitivity of the model to a disturbance applied on signal i is defined as [19]:

$$S_{ac(i)}^{auto}(i) = \frac{1}{N^{test}} \sum_{k=1}^{N^{test}} \left| \frac{\hat{X}_{nc}^{test-ac(i)}(k,i) - \hat{X}_{nc}^{test-nc}(k,i)}{X^{test-ac(i)}(k,i) - X^{test-nc}(k,i)} \right| \quad (3)$$

This metric measures the ability of the model to provide the same reconstructions in the two cases of disturbed or undisturbed signal i. In this respect, notice that a model characterized by a very low accuracy (high MSE) and very high robustness (small $S_{\text{ac}(i)}^{\text{auto}}(i)$) is not satisfactory for condition monitoring since it still provides signal reconstructions very different from signal values in normal plant operation.

The accuracy in the reconstruction of the disturbed signal i is defined as:

$$A_{ac(i)}^{auto}(i) = \frac{1}{N^{test}} \sum_{k=1}^{N^{test}} (\hat{X}_{nc}^{test-ac(i)}(k,i) - X^{test-nc}(k,i))^2 \quad (4)$$

This metric measures the mismatch between signal reconstructions and signal values in normal plant operation. However, since it does not consider either the difference between the reconstructions in the cases of disturbed and undisturbed signals, or the magnitude of the signal deviation

($X^{\text{test-ac}(i)}(k,i)$- $X^{\text{test-nc}}(k,i)$), it cannot be directly interpreted as a measure of model robustness.

Again, these two metrics actually measure errors and, thus, low values are desired.

Global robustness measures $\overline{S_{\text{ac}}^{\text{auto}}}$ and $\overline{A_{\text{ac}}^{\text{auto}}}$ are then obtained by applying a disturbance to all the signals, computing the robustness $S_{\text{ac}(i)}^{\text{auto}}(i)$ and $A_{\text{ac}(i)}^{\text{auto}}(i)$ and taking, respectively, the mean values:

$$\overline{S_{ac}^{auto}} = \frac{\sum_{i=1}^{n} S_{ac(i)}^{auto}(i)}{n} \tag{5}$$

$$\overline{A_{ac}^{auto}} = \frac{\sum_{i=1}^{n} A_{ac(i)}^{auto}(i)}{n} \tag{6}$$

3.1. Cross-validation procedure for the estimation of the performance metrics

In order to accurately estimate the values of the performance metrics on test sets of signals values not previously used in the model development, a cross-validation procedure can be adopted [15,16,24]. In particular, in the application that follows, the so called "*K*-fold" cross-validation error estimate is used to compare the performances [29]. The original dataset is randomly partitioned into K = 10 blocks of equal size. One of these blocks is used as validation data subset for the evaluation of the performance metrics of interest, and the remaining 9 blocks are combined together to constitute the training data subset. The cross-validation process is then repeated 10 times (the 10-folds), each time using a different block as validation set.

4. Genetic Algorithms

Genetic Algorithms (GAs) are optimization methods aimed at finding the global optimum of a set of real objective functions, $F \equiv \{f(\cdot)\}$, of one or more decision variables, $U \equiv \{u\}$, possibly subject to various linear or non linear constraints. Their main properties are that the search is conducted i) using a population of multiple solution points or candidates, ii) using operations inspired by the evolution of species, such as breeding and genetic mutation, iii) using probabilistic operations, and iv) using only information on the objective or search function and not on its derivatives.

GAs owe their name to their operational similarities with the biological and behavioural phenomena of living beings [1,2,12,17,20]. Correspondingly, the terminology adopted in GAs contains many terms borrowed from biology, suitably redefined to fit the algorithmic context. Thus, GAs operate on a set of (artificial) chromosomes, which are strings of numbers, generally sequences of binary digits 0 and 1. If the objective function of the optimization has many arguments (typically called control factors or decision variables), each string is partitioned in as many substrings of assigned lengths, one for each argument and, correspondingly, we say that each chromosome is partitioned in (artificial) genes. The genes constitute the so called genotype of the chromosome and the substrings, when decoded in real numbers, constitute its phenotype. When the objective functions are evaluated in correspondence of a set of values of the control factors of a chromosome, its values are called the fitness of that chromosome. Thus, each chromosome gives rise to a trial solution to the problem at hand in terms of a set of values of its control factors.

The GA search is performed by constructing a sequence of populations of chromosomes, the individuals of each population being the children of those of the previous population and the parents of those of the successive population. The initial population is generated by randomly sampling the bits of all the strings. At each step, the new population is then obtained by manipulating the strings of the old population in order to arrive at a new population**,** hopefully characterized by increased mean fitness. This sequence continues until a termination criterion is reached. As for the natural selection, the string manipulation consists in selecting and mating pairs of chromosomes in order to groom chromosomes of the next population. This is done by repeatedly performing on the strings the four fundamental operations of reproduction, crossover, replacement and mutation, all based on random sampling: the parent selection step determines the individuals which participate in the reproduction phase; reproduction itself allows the exchange of already existing genes whereas mutation introduces new genetic material; the substitution defines the individuals for the next population. This way of proceeding enables to efficiently arrive at optimal or near-optimal solutions.

With regards to their performance, it is acknowledged that GAs take a more global view of the search space than many other optimization methods. The main advantages are i) fast convergence to near global optimum, ii) superior global searching capability in complicated search spaces, iii) applicability even when gradient information is not readily available.

In a multi-objective optimization problem, several possibly conflicting objective functions $f_z(\cdot)$, i = 1, 2,…, n_f , must be evaluated in correspondence of each decision variable vector U in the search space. The goal is to identify the solution vector U^* which gives rise to the best compromise among the various objective functions. In this work, we adopt a simple aggregation method to combine the objectives into a scalar fitness function.

4.1. Genetic Algorithms for grouping optimization

For the task of grouping optimization for condition monitoring, the fitness function and the chromosome structure must be specified. The fitness function must evaluate the performance of the groups-based condition monitoring models: in our work, we use the metrics *MSE*, $\overline{S_{ac}^{auto}}$, $\overline{A_{ac}^{auto}}$ proposed in Section 3. For the chromosomes we simply take vectors made of a number of positive integers equal to the number of monitored signals: the integer value of the j-th element of the vector indicates the group to which the j-th signal is assigned.

Notice that according to the proposed chromosome structure, it is impossible to exclude a signal from all the groups or to assign a signal to more than one group. Furthermore, if 1 and R are the minimum and maximum values of the j-th gene, r_j, the maximum number of groups which can be used by the GA is R.

5. Application

A real case study concerning 46 signals used to monitor the RCP of a French PWR is considered. The signals values have been measured every hour for a period of 11 consecutive months and concern four RCPs, each one on a different line of the primary circuit. Experts have selected a dataset containing only data referring to normal condition operations of the system and, in order to discard possible outliers, those patterns characterized by value outside the interval $\mu \pm 3\sigma$, with μ indicating the signal mean and σ the signal standard deviation.

The 46-dimensional selected patterns (5798) have been divided into a set X_M of 2798 patterns used to perform the GA optimization, i.e., to find the optimal grouping, and a validation set X_v of 3000 patterns used to validate the condition monitoring performance on different data. In this respect, a 10-fold cross validation has been performed on the patterns of X_v . Thus, X_v has been randomly partitioned into K = 10 blocks of equal size. One of these blocks, X_v^{test} is used for the evaluation of the performance metrics of interest, and the remaining 9 blocks are combined together to constitute the training data subset, X_v^{train} . The cross-validation process has been then repeated 10 times (the 10-folds) using a different block as test set each time.

The results achieved by the GA-based wrapper approach are compared to those achieved by considering a single group formed by all signals and by grouping the signals according to their correlation [3]. In particular, the correlation grouping has been obtained by assigning to the same group the signals with an absolute value of the correlation coefficient larger than a threshold here set to 0.8; in other words, each signal in a group has at least a correlation larger than 0.8 with one of the other signals in the group. Following this procedure, 4 groups have been identified, whereas the remaining 4 signals, characterized by a correlation coefficient lower than 0.8 with all other signals, have been put together in a fifth group of uncorrelated signals. Notice that the application of this grouping approach requires to arbitrary fix a correlation threshold and to decide how to group the signals uncorrelated with all the others. To these purposes, a trial and error approach has been used in [6]: different values of the threshold have been considered to find the groups and the value which leads to the best compromise between accurate and robust reconstructions has been selected. Threshold values higher than 0.8 result in more than five groups which give more accurate but less robust reconstruction, whereas lower threshold values result in less than five groups characterized by less accurate but more robust reconstructions.

5.1. GA optimization

First, two single-objective GA searches, hereafter indicated as GA(*MSE*) and GA($\overline{A_{ac}^{auto}}$), have been performed with the objective of minimizing the metrics *MSE* and $\overline{A_{ac}^{auto}}$, respectively. In both cases, three runs of the GA search have been performed, and the

achieved solution with the best fitness has been considered. The maximum number of groups, *R*, has been taken equal to 5 in accordance to the number of groups used by the correlation grouping. From the results of experimentation performed by the authors in the present paper and in previous works [27] the GA parameters have been set. In particular, a population of 100 chromosomes has been considered in order to ensure enough genetic diversity in the population; the GA search ends when 150 generations are reached; at each generation, a new population of chromosomes is created from the previous one by using the standard roulette selection according to which the probability of choosing an individual as parent is proportional to its rank; the probability of mutation of each bit of the individual in the population has been set to 0.01. Table 1 reports the parameters which are recognized to most affect the GA performance [27].

Table 1

GA run parameters

Population Size (number of chromosomes in the population)	100
Number of Generations (termination criterion)	150
Selection Function	Standard roulette
Mutation Probability	0.01
Gene possible values	[1, 2, 3, 4, 5]

With respect to the AAKR reconstruction models developed within the wrapper approach, in order to reduce the computational efforts, only 200 randomly selected patterns constitutes the training set, X_M^{train}, and 100 different patterns the test set, X_M^{test}, used to compute the fitness function. The bandwidth parameter has been fixed to the value of 0.3 for all the developed AAKR models in a compromise between a small bandwidth, which results in reconstruction of the test patterns based on few patterns of X_M^{train} with associated high weights, and a large bandwidth (large value of h), which results in a reconstruction based on many patterns of X_M^{train} with relatively low weights.

Once the optimal grouping has been identified by the GA search, an AAKR model for each group has been developed in order to compute the performance of the grouping. For the development of these final models, a more refined optimization of the bandwidth parameter has been performed. In particular the training set X_v^{train} has been divided into two subsets, one used to train the AAKR model, the other to identify the optimal value of the bandwidth. To this purpose, 10 trial values (0.1, 0.2, …, 1.0) of the bandwidth have been considered. The obtained results show that the optimal bandwidth values are close to 0.3 for most of the groups.

Table 2 reports the performances of the obtained best groups with respect to the metrics *MSE*, $\overline{A_{ac}^{auto}}$ and $\overline{S_{ac}^{auto}}$ (although this latter has not been considered as objective function to be optimized) on the dataset X_V not considered during the GAs search. For the computation of the metrics $\overline{A_{ac}^{auto}}$ and $\overline{S_{ac}^{auto}}$ abnormalities in the signal behavior have been simulated by adding a random noise to the signals measured during normal plant operations. In particular, it has been assumed that during a plant transient only one signal is altered with respect to its value in normal operation, and the related deviation has been taken proportional to a Gaussian noise with zero mean and standard deviation equal to the 10% of the signal standard deviation.

Table 2

Single objective groupings performance

	GA(*MSE*)	GA($\overline{A_{ac}^{auto}}$)	correlation
MSE	0.007±0.001	0.006±0.001	0.008±0.001
$\overline{S_{ac}^{auto}}$	0.437±0.005	0.430±0.008	0.309±0.003
$\overline{A_{ac}^{auto}}$	0.010±0.001	0.009±0.0005	0.010±0.001

Table 3

Single objective groupings performance on the data used for the search

	GA(*MSE*)	GA($\overline{A_{ac}^{auto}}$)
MSE	0.011	0.012
$\overline{S_{ac}^{auto}}$	0.368	0.342
$\overline{A_{ac}^{auto}}$	0.015	0.013

Notice that the best grouping found with respect to the metrics *MSE* and $\overline{A_{ac}^{auto}}$ is obtained by GA($\overline{A_{ac}^{auto}}$), although the performances obtained by the other groupings are similar. This is due to the fact that the two metrics consider the difference between the signal reconstruction and the signal value in case of normal plant behavior (Eqs. 1, 4), the only difference being the signal values given in input to the condition monitoring model which are disturbed in case of $\overline{A_{ac}^{auto}}$. Furthermore, since the GA search is performed on a different set of data than that used for the evaluation of the performances X_V and the search is performed

on a small set of data in order to reduce the computational time, GA(*MSE*) tends to find a grouping more performing than GA($\overline{A_{ac}^{auto}}$) on the few data, X_M^{test}, used for the search (Table 3) but slightly less performing on the different data of X_v^{test} (Table 2). On the other side, using the metric $\overline{A_{ac}^{auto}}$ as fitness function in the GA search allows to obtain a signal grouping which is able to achieve higher accuracy when applied to different data than those used for the search (Table 2).

With respect to the robustness, the two groupings found by GA(*MSE*) and GA($\overline{A_{ac}^{auto}}$) are remarkably less satisfactory than the correlation grouping. On the other side, a GA search based on parameters of Table 1 with the objective of minimizing $\overline{S_{ac}^{auto}}$ has found an optimal grouping characterized by very low accuracy (0.028). This is due to the fact that a small $\overline{S_{ac}^{auto}}$ value indicates only that the signal reconstructions, although very similar in case of disturbed or undisturbed input signal, are in both cases very different from the desired signal values in normal plant condition [6].

These results have motivated the use of a multi-objective optimization problem which considers both the accuracy (measured by $\overline{A_{ac}^{auto}}$) and the robustness (measured by $\overline{S_{ac}^{auto}}$) as objectives of the search. The two objectives are aggregated into a single scalar function f_{agg} as:

$$f_{agg} = \overline{S_{ac}^{auto}} \cdot \overline{A_{ac}^{auto}} \quad (7) \qquad (8)$$

Since the optimization problem of identifying the best group for each one of the 46 signals requires a search between $5^{46} \approx 10^{32}$ possible solutions, the opportunity of reducing the search space by decreasing the number of signals that must be reorganized by the GAs has been considered. To this purpose, it has been observed that the correlation grouping contains a group formed by 30 highly correlated temperature signals which provides a very accurate and robust reconstruction of 24 signals. Thus, it has been decided to adopt an hybrid approach which consists in keeping these 24 signals in a group and using the GA search to optimize the group assignment of the remaining 22 signals. According to this hybrid approach, the chromosome is formed by a vector of 22 elements (the signals which should be assigned to a group) which can take a value between 1 and 5 (the group to which the signal is assigned). Notice that one of the group to which the signals can be assigned is the group of the 24 signals identified by the correlation grouping. In this way the dimension of the search space is reduced from the $5^{\mathbf{46}}$ ($\approx 10^{32}$) possible combinations considered in the search to $5^{\mathbf{22}}$ ($\approx 10^{15}$).

The GA approach is effectively able to reduce the fitness function $((f_{agg}^{(1)} - f_{agg}^{(150)})/f_{agg}^{(1)} = 70\%$, being $f_{agg}^{(1)}$ and $f_{agg}^{(150)}$ the fitness values of the best individual at the first and last generations, respectively). As expected, owing to the training and testing of many reconstruction models, the wrapper GA search has required a large computation time (10.2 hours on an Intel Core 2 duo, 3.17 GHz, 2GB RAM).

The obtained grouping (hereafter named hybrid) has been compared to the grouping obtained by considering the correlation criteria (hereafter called correlation). The two groupings are very similar, given that they differ only for a single group of 4 correlated signals to which a signal with low correlation with all the others has been assigned by the GA search.

Table 4 reports the performances of the groupings when they are tested on the validation set within a 10-fold cross validation approach. The hybrid and correlation grouping provides remarkably more accurate reconstructions than the single group formed by all signals. Considering the robustness, notice that the small $\overline{S_{ac}^{auto}}$ value achieved by the group formed by all the signals indicates only that the signal reconstructions, although very similar in case of disturbed or undisturbed input signal, are in both cases very different from the desired signal values in normal plant condition [6]. Thus, also from the point of view of the robustness the hybrid and correlation groupings are more satisfactory than the group formed by all signals. The comparison between the hybrid and correlation groupings shows that the two groupings are very similar, although the hybrid grouping is slightly more robust (lower value of $\overline{S_{ac}^{auto}}$) but less accurate than the correlation grouping.

Table 4

Performances of groupings without power and pressure signals

	All	hybrid	correlation
number of groups	1	5	5
MSE	0.0468±0.0032	0.0059±0.0009	0.0057±0.0009
$\overline{S_{ac}^{auto}}$	0.1001±0.0014	0.2867±0.0040	0.3022±0.0028
$\overline{A_{ac}^{auto}}$	0.0663±0.0032	0.0079±0.0008	0.0079±0.0009

The obtained results lead to the conclusion that in this case study the correlation criterion allows to identify a satisfactory grouping both from the point of views of accuracy and robustness. Since the grouping by the correlation criterion is less time consuming than the GA wrapper approach, the former has been more efficient.

With respect to the accuracy, notice that the AAKR reconstructs a pattern as a weighted sum of the values of its neighbors (Eq. A1). Thus, a factor of primary importance to obtain an accurate reconstruction is the coverage of the input space by the training patterns: more dense are the patterns in the training space, more accurate is the reconstruction of a test pattern since it will be primarily based on patterns characterized by a small distance from the test pattern and thus very similar to it. The higher is the correlation, the higher tends to be the density of the training patterns in the input space, and higher the accuracy in the reconstruction of the test pattern.

With respect to the robustness of the reconstruction obtained by using the AAKR model, notice that the higher is the number of input signals in a group, the higher tends to be the robustness of the group. This is due to the fact that AAKR reconstructions are based on the distance of the test pattern from the training patterns, computed in a high-dimensional space. Thus, the variation of one signal value leads to a small variation of the multidimensional distances, and consequently to very similar reconstructions in the cases of disturbed and undisturbed signals.

Finally, the reason for which correlation and hybrid groupings assign a low correlated signal to two different groups have been investigated. Notice that the creation of this group of uncorrelated signals is due to the practical necessity of assigning these signals to a group but it is not justified by the correlation criterion itself. For this reason, it is not surprising that the GA search finds a solution in which one of these low correlated signals is assigned to another group.

5.2. Application to the detection of an abnormal conditions

Once the grouping of the signals has been identified, it is possible to develop the reconstruction models for the on-line detection of the abnormal conditions. In practice, for each of the five groups of signals identified by the hybrid grouping, a dedicated AAKR model has been built.

In this Section, in order to test the performance of the proposed hybrid grouping in the detection of abnormal conditions, some tests have been conducted.

First of all, the AAKR reconstructions achieved by using the hybrid grouping have been compared to those achieved by a single group formed by all signals in case of a sensor failure. To this purpose, a small linearly increasing drift has been simulated in the sensor measuring the temperature of the water flowing to the first seal of the pump in line 1 (hereafter referred to as signal 4a). The drift starts at t=101 h and reaches a maximum amplitude of 1.6% of the mean signal value at t=600 h (Figure 3). Figure 4 reports the residuals $r(4a) = x^{test-ac}(4a) - \hat{x}_{nc}^{test-ac}(4a)$ i.e. the difference between the abnormal condition measurement, $x^{test-ac}(4a)$, and the reconstruction $\hat{x}_{nc}^{test-ac}(4a)$, obtained by using a single group formed by all signals (top) and by using the group found by the hybrid grouping approach (bottom).

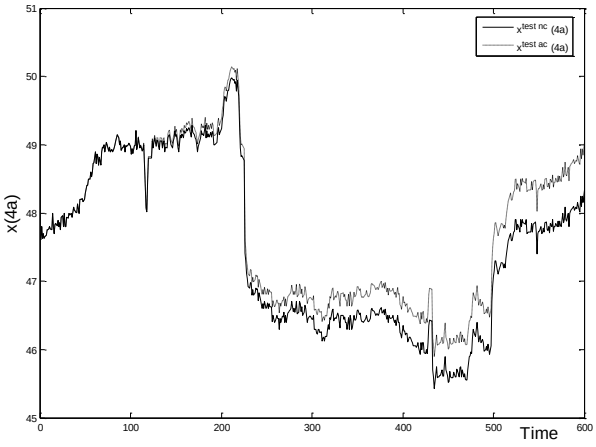

Fig. 3. Temperature of the water flowing to the first seal of the pump in line 1, $x^{test-nc}(4a)$, during 600 consecutive hours (continuous line) and its measurement obtained by a drifted sensor, $x^{test-ac}(4a)$ (dotted line).

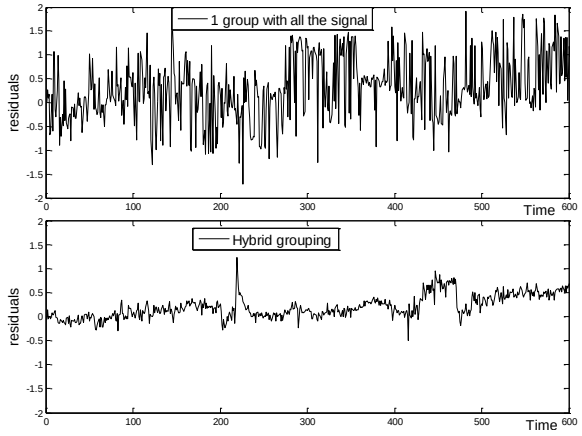

Fig. 4. residuals between the measured signal by the faulty sensor, $x^{test-ac}(4a)$, and its reconstruction $\hat{x}_{nc}^{test-ac}(4a)$ when a single group formed by all signals (top) and the hybrid grouping (bottom) are used for the signal reconstruction.

Based on the residuals of Figure 4, a decision about the health state of the sensor can be made. Residuals close to zero indicate normal conditions whereas remarkable deviations from zero for several consecutive time instants point out the onset of an abnormal condition. In the test case, the accuracy of the reconstruction achieved by the hybrid grouping before the beginning of the drift (t=101 h) is higher (residuals closer to zero) than that achieved by the single group of all signals: as a result, after the start of the drift at t=101 h, a clear tendency of the residuals to differ from zero is well detectable when the hybrid grouping is used, and less when the single group of all signals is used.

Notice that the execution times of the five AAKR models are very short (being the reconstruction performed in almost real time) and very similar to that of the single model formed by all signals.

In Figure 5 other failures of the same sensor have been simulated: a linearly increasing measurement noise of maximum intensity equal to 50% of the signal standard deviation (top), an offset of amplitude equal to 2% of the mean signal value (middle) and a stuck of the sensor (bottom). All sensor failures start at t=51 h. Figure 6 shows that in all cases the hybrid grouping residuals show a trend of deviation from zero, allowing detection of the anomalies occurred.

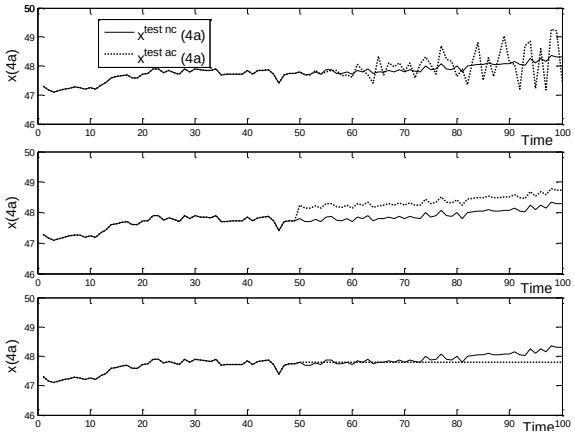

Fig. 5. Three different sensor failures: measurement noise increase (top), sensor offset (middle), sensor stuck (bottom).

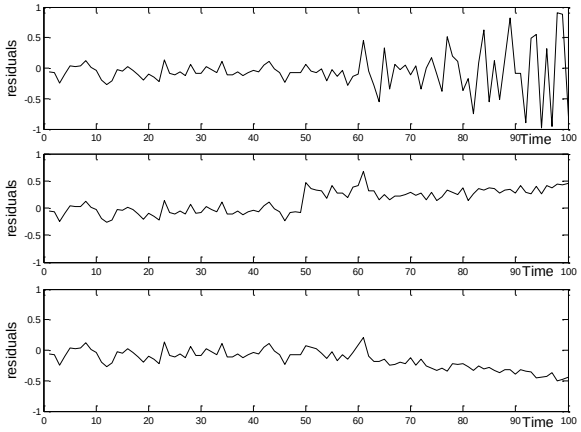

Fig. 6. Residuals obtained in the reconstruction of the three sensor failures of Figure 5.

In the last test, the performance of the grouping found by the hybrid approach has been verified with respect to the detection of a simulated deterioration of a RCP seal which leads to an increase of the seal flow. The abnormal condition (Figure 8, top, dotted line) has been simulated by adding a linear drift to the seal flow ($x(29a)$) collected in normal condition (Figure 8, top, continuous line). The simulated drift starts at t=51 h and reaches the maximum amplitude of 6.6% of the mean value of the signal at t=100 h. The residuals $r(29a) = x^{\text{test-ac}}(29a) - \hat{x}_{\text{nc}}^{\text{test-ac}}(29a)$, i.e. the difference between the abnormal condition measurement, $x^{\text{test-ac}}(29a)$, and the reconstruction $\hat{x}_{\text{nc}}^{\text{test-ac}}(29a)$ which is used for the abnormal condition detection, is reported in Figure 8, bottom. Notice that the residuals tend to be remarkably different from zero after t=70h, allowing the detection of the abnormal condition.

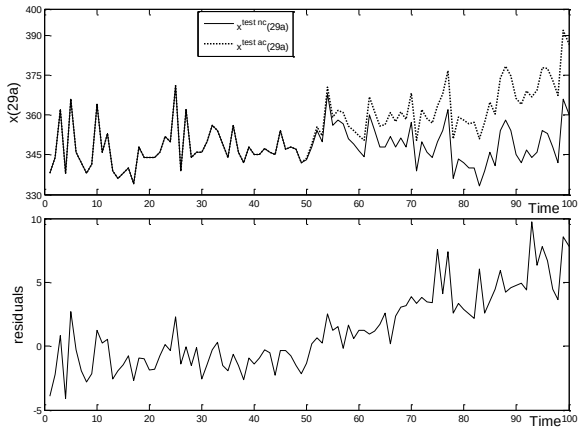

Figure 7. Top: time evolution of the first seal flow, $x^{test-nc}(29a)$, during 100 consecutive hours in normal condition (continuous line) and its evolution, $x^{test-ac}(29a)$, when a deterioration of the RCP seal is simulated (dotted line). Bottom: residuals between the measured signal, $x^{test-ac}(29a)$, and its reconstruction $\hat{x}_{nc}^{test-ac}(29a)$.

6. Conclusions

In this paper, we have proposed an approach to optimal grouping for condition monitoring. The approach is a wrapper, based on a GA search.

Since condition monitoring requires both accurate and robust reconstructions, we have taken a multi-objective point of view, reduced to a single-objective optimization problem by aggregation into a single scalar fitness function of two metrics of accuracy and robustness.

We have presented an application of the approach to the condition monitoring of RCP signals in which we have reduced the search space by a hybrid approach, in which the optimization of the group allocation considers only those signals which are not satisfactorily reconstructed by groups formed by highly correlated signals.

In the case study analyzed, the models built on the groups found by the GA search largely outperform that based on all signals. On the other hand, the groups obtained are very similar to those obtained based on signal correlation. This is due to the fact that:

- the AAKR modeling approach used for the signal reconstruction is more accurate on correlated signals;
- it has been shown that to be robust, the reconstruction model based on a group should also be accurate. For this, the signals in the group must be highly correlated. Furthermore, robustness of the reconstruction is related to the number of signals in the group: bigger groups lead to more robust AAKR reconstructions. In the considered case study, the presence of a big group of highly correlated signals guarantees both the robustness and the accuracy in the reconstruction of the signals.

In the case study analyzed, the proposed approach is computationally more demanding than the grouping based on signal correlation; however, it is expected to provide more robust reconstruction performances as the complexity of the problem increases. This is because for driving the search to the best grouping, the hybrid approach directly considers the performance of the condition monitoring model, whereas the correlation grouping is based on the correlation which is considered favorable for condition monitoring. Indeed, groups of highly correlated signals can provide accurate AAKR reconstructions, but the robustness of the reconstructions is more related to the dimension (number of signals) of the groups than to the correlation of the signals in the group.

In practice, a good compromise needs to be sought between the two conflicting objectives of low computational costs and robustness of the reconstruction by the models based on the groups: this depends on the application.

Furthermore, the high computational cost of the proposed approach occurs only in the phase of grouping the signals, which is performed only once, off-line. Then, a dedicated AAKR model is built for each group of signals, and the AAKR models can be executed in almost real time, which allows their use for online fault detection.

Future research work will focus on approaches of Multi-Objective Genetic Algorithms (MOGA) optimization in the framework of Pareto analysis, for finding a set of optimal solutions characterized by different compromises between accuracy and robustness. The results thereby obtained might differ from those of correlation grouping, particularly in very large search spaces. We will also consider other advanced optimization algorithms, such as Particle Swarm Optimization and Differential Evolution, for the search of the optimal grouping with reduced computational expenses.

Acknowledgements

This research has been carried out under contract C935C50010 funded by EdF R&D. Also, many thanks go to the reviewers for their constructive comments, which have allowed improving the paper.

Appendix A: the Auto-Associative Kernel Regression (AAKR) method

Let $X^{obs\text{-}nc}$ be a matrix of observed data whose generic element $X^{obs\text{-}nc}(k,j)$ represents the k-th time observation, k=1,…,N, of the j-th measured signal, j=1,…,q, taken during normal plant condition. The basic idea of the method is to reconstruct the signal values in case of normal condition, $\vec{\hat{x}}_{nc}$, given a current signal measurement vector, $\vec{x}^{obs}$ = ($x^{obs}(1)$,…,$x^{obs}(n)$), as a weighted sum of the observations in $X^{obs\text{-}nc}$. Thus, $\hat{x}_{nc}(j)$, the reconstruction of $x^{obs}(j)$, the j-th component of $\vec{x}^{obs}$, is given by:

$$\hat{x}_{nc}(j) = \frac{\sum_{k=1}^{N} w(k) \cdot X^{obs-nc}(k,j)}{\sum_{k=1}^{N} w(k)} \qquad \text{(A1)}$$

The weights $w(k)$ are similarity measures obtained by computing the Euclidean distance between the current sensor measurements $\vec{x}^{obs}$ and the k-th observation of $X^{obs\text{-}nc}$:

$$d^2(k) = \sum_{j=1}^{n} (x^{obs}(j) - X^{obs-nc}(j,k))^2 \qquad \text{(A2)}$$

and inserting it in the Gaussian kernel:

$$w(k) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{d^2(k)}{2h^2}} \qquad \text{(A3)}$$

where the signal h defines the Gaussian bandwidth.

In order to provide in (Eq. A2) a common scale across the different signals measuring different quantities, it is necessary to normalize their values. In the present work, the signal values are normalized according to:

$$x_n(j) = \frac{x(j) - \mu(j)}{\sigma(j)} \qquad \text{(A4)}$$

where $x(j)$ is a generic measurement of signal j and $\mu(j)$ and $\sigma(j)$ are the mean and standard deviation of the j-th signal in $X^{obs\text{-}nc}$:

$$\mu(j) = \frac{\sum_{k=1}^{N} x^{obs-nc}(k,j)}{N}$$

$$\sigma(j) = \sqrt{\frac{\sum_{k=1}^{N} \left(x^{obs-nc}(k,j) - \mu(j)\right)^2}{N}}$$

References

[1] H. Adeli and N.-T Cheng, '*Augmented Lagrangian Genetic Algorithm for Structural Optimization*', Journal of Aerospace Engineering, **7**, pp. 104-118, 1994.

[2] H. Adeli and N.-T Cheng, '*Concurrent Genetic Algorithms for Optimization of Large Structures*', Journal of Aerospace Engineering, **7:**3, pp. 276-296, 1994.

[3] H. Adeli and S. Kumar, '*Distributed Genetic Algorithms for Structural ptimization*', Journal of Aerospace Engineering, **8**:3, pp. 156-163, 1995.

[4] A. Al-Bazi and N. Dawood, '*Developing Crew Allocation System for Precast Industry Using Genetic Algorithms*', Computer-Aided Civil and Infrastructure Engineering, **25**, pp. 581-595, 2010.

[5] P. Baraldi, A. Cammi, F. Mangili and E. Zio, '*An ensemble approach to sensor fault detection and signal reconstruction for nuclear system control*', Annals of Nuclear Energy, **37**, pp. 778-790, 2010.

[6] P. Baraldi, R. Canesi, E. Zio, R. Seraoui and R. Chevalier, '*Signal Grouping for Condition Monitoring of Nuclear Power Plant Components*', Seventh American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies NPIC&HMIT 2010, Las Vegas, Nevada, November 7-11, 2010.

[7] P. Baraldi, E. Zio, G. Gola, D. Roverso and M. Hoffmann, '*Genetic Algorithms for Signal Grouping in Sensor Validation: a Comparison of the Filter and Wrapper Approaches*', Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, **222**, pp. 189-206, 2008.

[8] P. Baraldi, E. Zio, G. Gola, D. Roverso and M. Hoffmann, '*Robust nuclear signal reconstruction by a novel ensemble model aggregation procedure*', International Journal of Nuclear Knowledge Management, **4**, pp 34-41, 2010.

[9] P. Baraldi, G. Gola, E. Zio, D. Roverso and M. Hoffmann, '*A randomized model ensemble approach for reconstructing signals from faulty sensors*', Expert System With Application, **38** , pp. 9211-9224, 2011.

[10] P. Baraldi, E. Zio, G. Gola, D. Roverso and M. Hoffmann, '*Signal reconstruction by a GA-optimized ensemble of PCA models*', Nuclear Engineering and Design, **241**, pp.301-309, 2011.

[11] P. Baraldi, E. Zio, G. Gola, D. Roverso and M. Hoffmann, '*Two novel procedures for aggregating randomized model ensemble outcomes for robust signal reconstruction in nuclear power plants monitoring systems*', Annals of Nuclear Engineering, **38**, pp. 212-220, 2011.

[12] L. Chambers, '*Practical handbook of genetic algorithms: applications*' Vol. 1, CRC Press, 1995.

[13] R. Chevalier, D. Provost and R. Seraoui, *'Assessment of Statistical and Classification Models For Monitoring EDF's Assets*', Sixth American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies, Knoxville, USA, 2009.

[14] L. Dridi, M. Parizeau, A. Mailhot, J.P. Villeneuve, '*Using Evolutionary Optimisation Techniques for Scheduling Water Pipe Renewal Considering a Short Planning Horizon*', Computer-Aided Civil and Infrastructure Engineering, **23,** pp. 625-635, 2008.

[15] B. Efron, '*Estimating the error rate of a prediction rule: Improvement on cross-validation*', Journal of the American Statistical Association, **78**, pp. 316-331, 1983.

[16] B. Efron and R.J. Tibshirani, '*Improvements on cross-validation: The .632+ bootstrap method*', Journal of the American Statistical Association, **92**, pp. 548-560, 1995.

[17] D.E. Goldberg, '*Genetic algorithms in search, optimization, and machine learning*', Addison-Wesley Publ. Co., 1989.

[18] K.C. Gross and K.E. Kumenik, '*Sequential probability ratio test for nuclear power plant component surveillance*', Nuclear Technology **93**, pp. 131–137, 1991.

[19] J. W. Hines and D. R. Garvey, '*Development and Application of Fault Detectability Performance Metrics for Instrument Calibration Verification and Anomaly Detection*', Journal of Pattern Recognition Research, **1**, pp 2-15, 2006.

[20] J.H. Holland, '*Adaptation in natural and artificial systems: an introductory analysis with application to biology*', Control and Artificial Intelligence. MIT Press, 4-th edition, 1975.

[21] S.L. Hung and H. Adeli, '*A Parallel Genetic/Neural Network Learning Algorithm for MIMD Shared Memory Machines*', IEEE Transactions on Neural Networks, **5**:6, pp. 900-909, 1994.

[22] M.W. Kang, P. Schonfeld, N. Yang, '*Prescreening and Repairing in a Genetic Algorithm for Highway Alignment Optimization*', Computer-Aided Civil and Infrastructure Engineering, **24**, pp. 109-119, 2009.

[23] H. Kim and H. Adeli, '*Discrete Cost Optimization of Composite Floors using a Floating Point Genetic Algorithm*', Engineering Optimization, **33**:4, pp. 485-501, 2001.

[24] R. Kohavi, '*A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*', San Mateo, CA: Morgan Kaufman. pp. 1137–1143, 1995.

[25] R. Kohavi and G. John, '*Wrappers for feature subset selection*', Artificial Intelligence, **97**, pp. 273-324, 1997.

[26] H. Lee, E. Kim, M. Park, *'A genetic feature weighting scheme for pattern recognition*', Integrated Computer-Aided Engineering, 14, pp. 161–71, 2007.

[27] M. Marseguerra, E. Zio and S. Martorell, '*Basics of genetic algorithms optimization for RAMS applications*', Reliab Eng Syst Saf, **91**, pp 977-991, 2006.

[28] S. Mathakari, P. Gardoni, P. Agarwal, A. Raich and T. Haukaas, '*Reliability-based Optimal Design of Electrical Transmission Towers Using Multi-objective Genetic Algorithms*', Computer-Aided Civil and Infrastructure Engineering, **22**, pp. 282-292, 2007.

[29] R. Polikar, '*Bootstrap-Inspired Techniques in Computational Intelligence*', IEEE signal processing magazine, **59**, pp. 59-72, 2007.

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

[31] D. Roverso, M. Hoffmann, E. Zio, P. Baraldi and G. Gola, '*Solutions for plant-wide on-line calibration monitoring*', Proc. ESREL 2007, Stavanger, Norway, **1**, pp.827–832, 2007.

[32] K. Sarma, and H. Adeli, '*Life-Cycle Cost Optimization of Steel Structures*', International Journal for Numerical Methods in Engineering, **55**:2, pp. 1451-1462, 2002.

[33] J. F. Smith and **T. H. Nguyen,** '*Autonomous and cooperative robotic behavior based on fuzzy logic and genetic programming',* Integrated Computer-Aided Engineering, **14**:2, pp. 141-159, 2007.

[34] F. Teklu, A. Sumalee and D. Watling, '*A Genetic Algorithm for Optimizing Traffic Control Signals Considering Routing*', Computer-Aided Civil and Infrastructure Engineering, **22**, pp. 31-43, 2007.

[35] A. Tucker, S. Swift, and X. Liu, '*Variable Grouping in Multivariate Time Series Via Correlation*', IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, **31**, 2001.

[36] E. Zio, P. Baraldi and N. Pedroni, '*Selecting Features for Nuclear Transients Classification by means of Genetic Algorithms*', IEEE Transactions on Nuclear Science, **53**, pp. 1479-1493, 2006.