Optimization of the features extraction method in cyber physical systems of monitoring energy infrastructure facilities

Vladislav Kats[1,*] and *Andrey* Volkov[1]

[1] Moscow State University of Civil Engineering, Yaroslavskoe shosse, 26, 129337, Russia

Abstract. The study presents an optimization method of sliding windows parameters selection in the features extraction procedure in cyber physical systems during data processing of the technical condition of energy and electric infrastructure facilities. The method is based on utilizing genetic algorithm and objective functions related to the quality of the classification of the defects by hazard degree. Proposed method was verified on the experimental data acquired from the technical condition monitoring of the vertical oil tanks. The results obtained from the experiment confirm the proposal that this method can be applied for effectively solving problems related to features extraction in monitoring, risk management and classification of maintenance defects in the power and electrical engineering facilities.

1 Introduction

Evolution of the defects in energy infrastructure facilities can lead to catastrophic consequences and huge economic losses. However, timely detection of the evolving defect in the structure of controlling object allows to prevent multiple risks related to even temporary decommissioning. It is an important issue for a variety of facilities of power and electrical engineering domain such as wind turbines and diesel engines [1, 2].

Accordingly, the development of cyber physical systems of diagnostic monitoring [3], that allows detecting defects, identifying its hazard degree and forecasting their evolution, becomes a relevant problem.

When designing monitoring system, engineers often use sensors that record monitoring data representing acoustic emission (AE) signals time series [1].

In order to implement effective defects classification and to forecast their evolution, it is necessary to acquire the extended representation of AE signal time series. Meanwhile, so called features extraction methods are usually adopted [4]. Features extraction methods map AE time series to diagnostic parameters matrix, that can be further used for the classification of the defects by their hazard degree. AE sensors data has an important peculiarity: they are usually being recorded with high sampling frequency no less than 1-2MHz that leads to the formation of a huge amount of monitoring data [5]. Consequently, providing data compression during features extraction is particularly important. In addition, it is important to provide a sensitivity of these features to AE time series' local structure evolution at occurrence of a small duration single impulse.

An issue of features extraction frequently appears in problems of the desired signal recognition against the operational, technological and other types of noise of different nature. It is shown in the work [6] that an effective method of informative features extraction from AE time series exists. It is based on consequent calculations of waveform statistical parameters on both the minor scale corresponding to analyzing windows and on the major scale corresponding to texture sliding windows.

A set of diagnostic features are usually selected on the basis of the application domain and described in work [7]. However, this approach of feature extraction from AE time series leads to a major challenge. In order to apply this approach it is necessary to calculate features in sliding windows with overlays. Meanwhile, the selection of the window width and the overlay ratio generally may be quite subjective. Therefore, it does not allow to acquire clear representation of AE time series to the matrix of diagnostic features, which has a negative impact on further defects classification by the hazard degree. Moreover, it is widely spread that feature extraction algorithms suppose utilizing multiple setup parameters which selection depends on human factor and directly influences on the quality and on the level of confidence to results of the forecast and the classification of the defect by hazard degree.

Authors of the paper [7] recommend to determine the width of an analyzing window from the following considerations. First, the width of the window should be enough small to provide stationarity of the noisy signal and its spectrum within the window bounds. Second, the width of the window should be large enough for providing calculation representability in the time domain and for required spectral resolution in the frequency domain. The width of the texture window should be small enough to distinguish noise and signal components but it should be large enough to identify signal

* Corresponding author: vladk_94@mail.ru

components of different hazard classes that appears during the long-lasting evolution of the defects.

It is significant that given recommendations allows to define only the range of possible values of both windows widths on minor and major time scales and their overlays ratios. Meanwhile, experimental data shows that the features extraction results considerably depend on the values of sliding windows and overlays [8].

In current paper, we propose a method based on genetic algorithm of the selection of optimal window width and overlay ratio, taking into account the specific nature of the controlling facility and excluding the impact of a human factor [9].

The description of the proposed method is given in section two. Section 3 provides the verification of proposed method on experimental data.

2 Method

A genetic algorithm is a heuristic search algorithm used to solve optimization and modelling problems by sequentially selecting, recombining, and varying the desired parameters, using mechanisms similar to those by which biological evolution is performed [9]. A distinguishing feature of the genetic algorithm is an accent to utilizing of a crossover operation, which makes recombination of solutions (candidates) which role is similar to the crossover operation in nature. Genetic algorithms are also utilized in power engineering industry. In paper [10], the genetic algorithm was successfully applied for the fault diagnosis of smart grids. The advantages of genetic algorithms:

- A lot of parameters that allow to effectively utilize heuristics.
- Effective parallel computation.
- It works no worse than the random search.

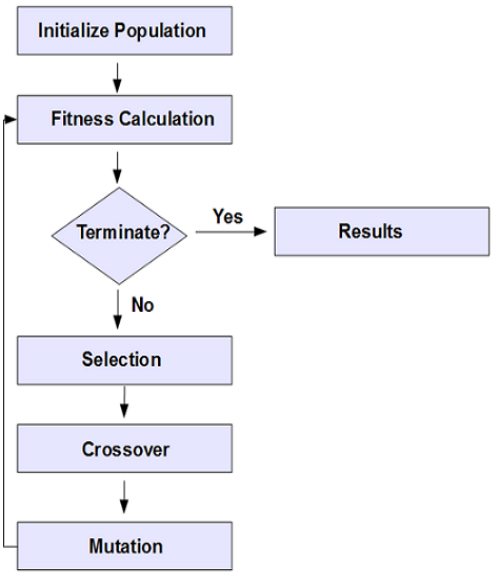

Fig. 1 Functional scheme of genetic algorithm [9]

An important factor that defines the effective performance of genetic algorithm is an appropriate selection of the fitness function. However, the principal purpose of the problem of optimal window and overlay size selection is to improve quality of the defects classification by hazard degree within the context of features extraction. Accordingly, it is necessary to define a fitness function as function characterizing the quality of the classification.

There are several metrics of classification quality. Each of them has its benefits and drawbacks. In current study, we considered a following set of classification quality metrics: Rand Index, mutual information coefficient, silhouette score, Davies-Bouldin Index, homogeneity score, Fowlkes-Mallows score and Calinski-Harabasz Index. After the preliminary experimental results, we selected three following metrics that provide the maximal informative content of the features along with the convergence speed: Rand Index, mutual information coefficient and Davies-Bouldin Index.

Rand Index (RI). The Rand Index computes a similarity measure between two classes by considering all pairs of samples and counting pairs that are assigned in the same or different classes [11].

$$\mathrm{RI} = \frac{a+b}{C_2^{n_{samples}}} \quad (1)$$

where a the number of pairs of elements that are in the same set in C and in the same set in K, b is the number of pairs of elements that are in different sets in C and in different sets in K. The advantages of RI is that:

- Random assignments have a RI coefficient close to 0.0 for any value of C.
- It has range within [0; 1]; values that are close to zero indicate on dissimilar classes, while similar classes have positive RI close to 1.0 [9].

Mutual Information coefficient. The mutual information calculates between to classification results as follows:

$$\mathrm{MI}(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)\log\left(\frac{P(i,j)}{P(i)P'(j)}\right), \quad (2)$$

where U – classification result after window size and overlay selection, V – the hazard degree estimated by other methods. It has upper bound of 1: values close to zero indicate that results of classification are different, while values close to one indicates that they are similar (with or without permutation). The major drawback of MI coefficient is that it requires the knowledge of the ground truth classification results [12].

Davies-Bouldin Index. The estimate of how well the clustering has been done is made by using quantities and features inherent to the dataset [12].

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\frac{s_i + s_j}{d_{ij}}, \quad (3)$$

where s_i is the average distance between each point of cluster and the centroid of that cluster – also known as cluster diameter. d_i is the distance between cluster centroids. The drawback of the metric is that the usage

of centroid distance limits the distance metric to Euclidean space.

3 Results

Verification of the proposed method of the optimal window size and the overlay ratio was carried out on the basis of features extraction in monitoring the technical condition of the energy infrastructure facility (oil tank). Figure 1 presents fragments of experimental AE time series obtained after the non-destructive testing of the object, where a, b and c letters correspond to defects of different hazard degree located in the tested object.

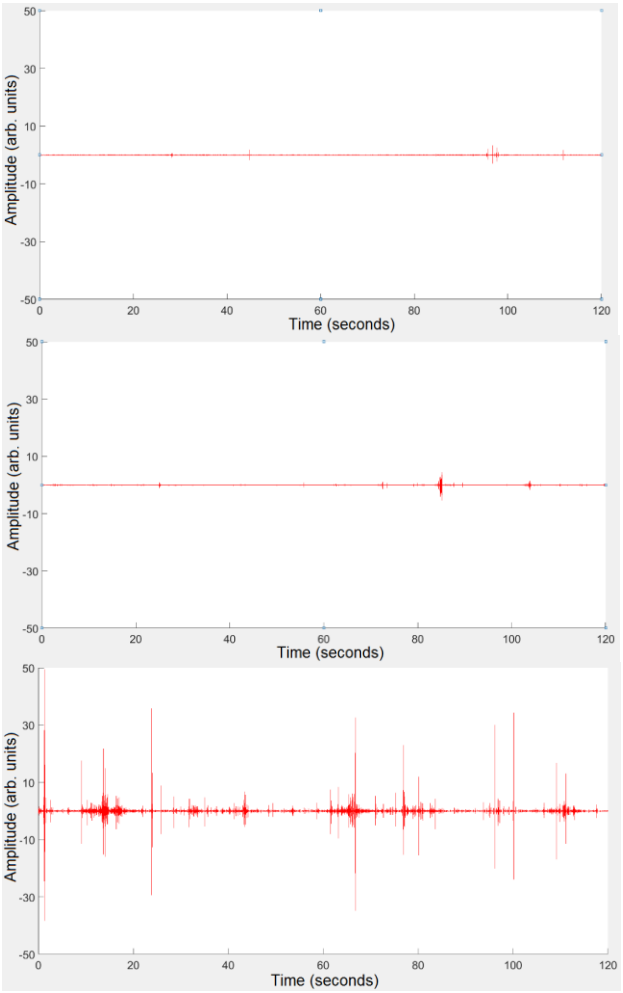

Fig 2. Fragments of the noisy AE time series. The inspected object: corner weld joint of the building construction (vertical steel oil tank #3. Volume – 1000 cubic meters, fuel-handling facility, AO "NTEK"). Sampling frequency – 2.5MHz. a) The defect of the 1st hazard class. b) The defect of the 2nd hazard class. c) The defect of the 3rd hazard class.

In the current paper, we utilize a series of diagnostic features forming a set of statistical parameters. These parameters describe a waveform of AE signals in time and frequency domain: spectral centroid, spectral spread spectral flux, spectral roll-off, entropy and power [7]. They were calculated for sliding windows of different window width upon minor and major time scales, which were described in section one.

Figure 3 shows the convergence of the genetic algorithm to optimal values of window sizes and overlay ratios while using the mutual information coefficient as a fitness function. It is clearly shown, that the algorithm converges already on 20-th iteration: best values of fitness function (bottom plot) begin fluctuations around optimal value and, meanwhile, mean values (top plot) decrease monotonously.

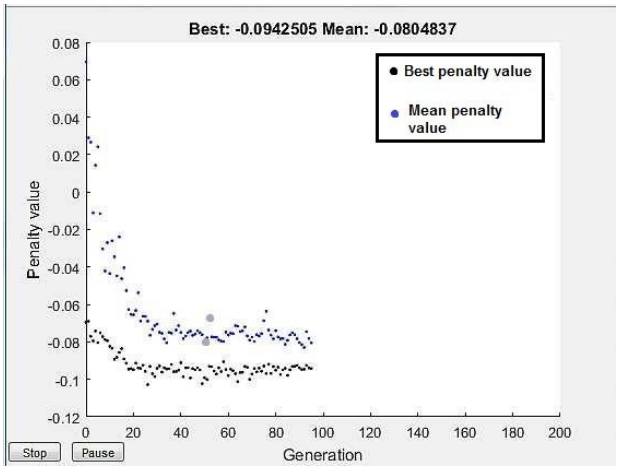

Fig. 3 The algorithm convergence to the optimal values of window width and overlay ratio

Table 1 presents results of comparison of the algorithm convergence speed when using different metrics of fitness function. It follows from the table 1 that the maximum speed of convergence was reached by utilizing MI coefficient as fitness function.

Table 1. The convergence speed of the proposed method in the comparison of fitness function selection.

Fitness function	Number of iterations
Mutual Information	20
Rand Index	33
Davies-Bouldin	67

A following set of optimal parameters of analyzing and texture windows and their overlays was obtained as a result for MI metric: analyzing window width – 0.32ms, analyzing overlay ratio – 0.19ms, texture window width – 4.33ms, texture overlay ratio – 3.21ms.

The verification of the informative content of features extraction results for selected optimal values of window width and overlay ratios supposes utilizing machine-learning techniques in the multiple dimension features space. In current paper, we adopt "Stochastic neighbor embedding" (SNE) method for the reduction of the feature space and for the visualization multivariate parameters [13]. The main feature of SNE method is maintaining the local structure and probabilistic properties of the source data in lower dimension space that allows to identify the clustering of the data in two or three features space. Given peculiarity was used in

current paper for analysis of futures extraction results upon the optimized windows and overlays sizes.

Figure 4 presents results of adopting SNE method to 48-dimensional diagnostic matrix calculated on the previous step. A preliminary standardization of the matrix has been conducted before applying SNE method. Barnes-Hut version of SNE method was employed for the sake of calculation performance [13]. Each point on Fig. 4 represents a feature vector corresponding given texture window in three-dimensional SNE components space. It follows from Fig. 4 that a distinct grouping of points that correspond to the certain set of texture windows takes place. The number of these groups corresponds with the number of different hazard classes of defects in structure of oil tank (Fig. 2). Therefore, it can be assumed that proposed set of statistical parameters may be employed as an informative feature in the feature identifying data clustering and recognition processes in cyber physical systems of monitoring energy facilities.

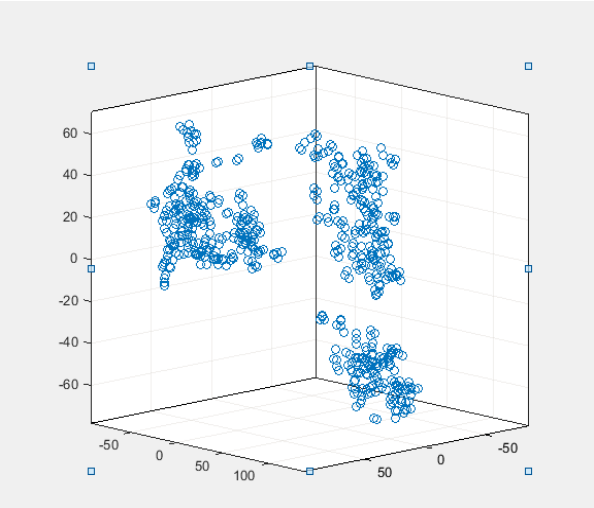

Fig. 4 Averaged statistical features in the space of three SNE components (optimal values of window sizes and overlay ratios have been used).

4 Conclusions

In current paper we proposed a genetic algorithm based method of the selection of optimal windows sizes and overlays ratios in the context of a features extraction problem.

It has been shown that the algorithm convergence speed is the largest when utilizing the fitness function built upon the mutual information coefficient.

It has been established that results obtained can be applied for effectively solving problems related to monitoring, risk management and classification of maintenance defects in the power and electrical engineering facilities, and particularly can be used in cyber physical systems utilizing non-destructive testing data.

Future work is related to the development of methods of the defects classification by their hazard degree on the basis of selected features obtained in current study.

Current study was performed under the financial support of RF President's grant #NSh-3492.2018.8.

References

1. J. Van Dam, L. Bond, *Proc. SPIE* **9439**, 3 (2015)
2. P. Tavner, IET Electric Power Applications, **2,** 4, pp. 215–247 (2008)
3. A. Volkov and E. Nasonov, Sistemotehnika stroitelstva, kiberfizicheskiye stroitel'niye sistemi, pp. 184-188 (2018)
4. L. Calabrese, G. Campanella, E. Proverbio, Construction and Building Materials, **34**, pp. 362-371 (2012)
5. D. Aggelis, *Innovative AE and NDT Techniques for On-Site Measurement of Concrete and Masonry Structures*, **20**, pp. 69-88 (2016)
6. G. Tzanetakis, P. Cook, *IEEE Transactions on Speech and Audio Processing*, **10**, pp. 293-302 (2002)
7. Fig. 1. «Genetic Algorithm», 08.10.2019, URL : https://apacheignite.readme.io/docs/genetic-algorithms
8. T. Giannkopoulos, A. Pikrakis, *Introduction to Audio Analysis, A MATLAB Approach* (2014)
9. V. Barat, Chernov D., Elizarov S. Russian journal of nondestructive testing, **52,** 6, pp. 347-356 (2016)
10. Zh. Peng, L. Na, Qu Bo-yang, International Journal of Smart Grid and Clean Energy, **7**, pp. 170-179 (2018)
11. W. Rand, J. of the American Statistical Association. **66**, 846–850 (1971)
12. E. Archer, I. Park, J. Pillow, Entropy, **15**, 12, pp. 1738–1755 (2013)
13. L. van der Maaten, G. Hinton, J. of Machine Learning Research, **9**, pp. 2579-2605 (2008)