Feature Selection - Extraction Methods based on PCA and Mutual Information to improve Damage Detection problem in Offshore Wind Turbines

Ekhi Zugasti^{1,a}, Luis E. Mujica^{2,b}, Javier Anduaga¹, Fernando Martinez¹.

¹IK4-Ikerlan, Sensors Department, Spain

²Department of Applied Mathematics III, Universitat Politècnica de Catalunya, Spain

^aezugasti@ikerlan.es, ^bluis.eduardo.mujica@upc.edu

Keywords: Damage Detection, Structural Health Monitoring, Feature Selection, Feature Extraction, Offshore

Abstract. Damage detection problem in Structural Health Monitoring (SHM) is widely studied by many researchers, therefore lots of damage detection algorithms can be found in the literature. Feature Selection / Extraction methods are essential in the accuracy of these algorithms, they provide the suitable data to be used. The main goal of this work is to improve the input data to be the most representative for the damage detection problem. This is done using different Feature Selection / Extraction methods (PCA, UmRMR, and a combination of both). After taking the representative features, the results are tested using a damage detection method; the NullSpace in this case. The data has been collected from a Laboratory Offshore tower model. The different results are compared (different pre-processing vs Raw data) and these show how the correct preselection of the data can improve damage detection.

Introduction

Structural Health Monitoring (SHM) aims to give, at every moment during the life of a structure, a diagnosis of the "state" of the composing materials, different parts and, the full assembly of these parts which constitute the structure as a whole. The state of the structure must remain in the domain specified in the design, although this can be altered by normal aging due to usage, by the action of the environment or by accidental events. Damage detection is the first level of a SHM system and it is very important to avoid either human or economical disasters. If a damage can be detected in the moment that it is created, some actions could be conducted before a disaster happens. In the literature, a lot of methodologies for damage detection can be found, among them vibration based methods are the most used as shown in [1].

The objective of this work is to find a strategy to preprocess the original data (dynamical responses of a structure) in order to take into account the most suitable information and therefore, to improve the accuracy of the damage detection technique. To test the Feature Selection / Extraction methods proposed in this work, the NullSpace damage detection method described in the paper [2] is used. Please refer to that paper for more details on the specific damage detection algorithm.

The paper is organized as follows: Firstly, the Feature Extraction and Feature Selection methods are introduced. Afterward, the integration of these methods with the damage detection technique is illustrated. Subsequently, the test structure is presented and finally, results and conclusions are discussed.

Feature Extraction based on Principal Component Analysis

When an algorithm has to process a huge input data, or when this input data is suspected to be redundant, this data will be transformed into a representation set of features. Transforming the input data into the set of features is called Feature Extraction. If these extracted features are carefully chosen, it

is expected that they obtain the relevant information from the original input data in order to perform the desired task [3]. In this work, Principal Component Analysis (PCA) is used as Feature Extraction method [4]. Although it can be found another methods with similar results such as: Independent Component Analysis (ICA) [5], Nonlinear Dimensionality Reduction (NLDR) [6] or Partial least squares regression (PLS regression) [7].

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. Considering the original data arranged in a data matrix $\mathbf{X}$, with zero empirical mean, where each one of the m rows represents a different repetition of the experiment, and each one of the n columns gives a particular kind of datum. The goal of PCA is to find a $m \times n$ linear transformation matrix $\mathbf{P}$, which is used to transform the original data matrix $\mathbf{X}$ into the form:

$$\mathbf{T} = \mathbf{XP}, \quad (1)$$

where $\mathbf{P}$ can be determined by means of the singular value decomposition (SVD) of the covariance matrix. The new coordinates in PCA are defined by the eigenvectors and eigenvalues of the covariance matrix as follows:

$$\mathbf{C}_X = \frac{1}{m-1} \mathbf{X}^T \mathbf{X}, \quad \mathbf{C}_X \mathbf{P} = \mathbf{P} \Lambda. \quad (2)$$

where the eigenvectors of $\mathbf{C}_X$ are the columns of $\mathbf{P}$, and the eigenvalues are the diagonal terms of Λ (the off-diagonal terms are zero). The eigenvectors p_j forming the transformation matrix $\mathbf{P}$ are sorted according to the eigenvalues by descending order and they are called the principal components of the data set. The eigenvector with the highest eigenvalue represents the most important pattern in the data with the largest quantity of information.

Geometrically, the transformed data matrix $\mathbf{T}$ (Score Matrix) is the projection of the original data $\mathbf{X}$ over the direction of the principal components $\mathbf{P}$. PCA also seeks to reduce the dimensionality of the data set by choosing only a reduced number r of principal components ($r < n$). Now, the $\mathbf{T}$ given by the reduced matrix $\mathbf{P}$, is smaller than the original data ($\mathbf{X}$).

Feature Selection based on Mutual Information

In machine learning and statistics, Feature Selection, also known as variable selection, is the process of selecting a subset of the original features (or variables) which are the more appropriate to use in the model construction. This is based on the fact that the original data contains redundant or irrelevant features. Redundant features are those which provide no more information than the other features, while irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of Feature Extraction. Feature Extraction creates new features from functions of the original features, whereas Feature Selection returns a subset of the features [8,9]. Feature selection techniques provide three main benefits when constructing models: improving model interpretability, shorter training times and enhanced generalization by reducing Overfitting [10]. Lots of these methods use Mutual Information (MI) as a measure of similarity [11], for instance: Random Sampling (RS) [12], Mutual Information Maximization (MIM) [12], Mutual Information Feature Selection (MIFS) [13] or minimum Redundancy - Maximum Relevance (mRMR) [14]. A large list of the Feature Selection methods can be found on [15]. In this work, Unsupervised minimum Redundancy - Maximum Relevance (UmRMR) method is implemented [16].

The goal of UmRMR is to find those variables which have minimal redundancy and maximal relevance, by using Mutual Information to measure the level of similarity between features and Entropy to gauge the level of Information of each feature. The relevance of a feature is defined as the average

Mutual Information to the whole feature set as follows:

$$Rel(x_i) = \frac{1}{n} \sum_{j=1}^n I(x_i; x_j) = \frac{1}{n} \left(H(x_i) + \sum_{1 \leq j \leq n, j \neq i} I(x_i, x_j) \right), \quad (3)$$

where $H(x_i)$ indicates the information content included in feature x_i (Entropy of x_i), which should be large, and the other term is the amount of information content included in all the other features due to the knowledge of x_i (Mutual Information between x_i and the rest of variables), which should be also large. If we select feature x_i which has the maximal $Rel(x_i)$, then it can lead to the loss of information to the least extent.

In order to define redundancy, Xu [16] defines first the conditional relevance (eq. 4), to know what is the information gain of selecting y_i if we have already selected x_i . So, the redundancy (eq. 5) is defined as the difference between the relevance of y_j and the conditional relevance of y_j if x_i is previously selected.

$$Rel(y_i|x_i) = \frac{H(y_i|x_i)}{H(y_i)} Rel(y_i), \quad (4)$$

$$Red(x_i; y_i) = Rel(y_i) - Rel(y_i|x_i). \quad (5)$$

To define which is the variable with minimal redundancy and maximal relevance, a sequential forward search is performed to rank the features according to the following equation:

$$l_m = arg \max_{1 \leq i \leq n, x_i \in U} \left\{ Rel(x_i) - \frac{1}{m-1} \sum_{y_j \in S_{m-1}} Red(x_i; y_j) \right\} \quad (6)$$

Data preprocessing for damage detection

In this paper, three types of data preprocessing methods are implemented: the first one for Feature Extraction, the second one for Feature Selection and, the last one for Feature Selection-Extraction, which combines the former.

Feature Extraction methodology. As was previously explained, this method is based on PCA. The schematic view of the complete methodology for damage detection is depicted in Fig. 1a. The damage detection algorithm (NullSpace) is used as a tool to compare different preprocessing methods. For more information about this damage detection method see [2]. The preprocessing is performed as follows: First of all, a dataset belongs to the healthy structure is used to calculate the Principal Components (PCs). Afterward, the dataset gathered from the current structure (learning and detection datasets) are projected into the choose PCs. The projected data is used as the input to the NullSpace damage detection method.

Feature Selection methodology. This method which is based on UmRMR is depicted in Fig. 1b. As the previous methodology, the NullSpace [2] is used as a tool. This preprocessing is also performed in two steps: Firstly, it is needed to find which sensors are the most representative for the healthy state classification. In other words, UmRMR is applied to all the datasets obtained from the healthy structure. The algorithm gives the order of the sensors, from the most important to the least one. The number of selected sensor depends on the score of each sensor. Finally, the channel selection is performed to select the sensors of the incoming datasets (learning and detection datasets). Therefore, the selected sensors are used as input for the damage detection method.

Feature Selection-Extraction methodology. This preprocessing method uses first the UmRMR Feature selection method, and later the PCA Feature Extraction method. In Fig. 1c the schematic view of the methodology is shown. The preprocessing needs two steps as well, the first one for selecting

Fig. 1: Data preprocessing methods for damage detection

the sensors and extracting the features and, the second one to apply the results of the first step to the incoming datasets (learning and detection datasets). In the first step, the first part is the same as the one presented in the previous subsection, the most informative sensors are found using UmRMR. After that, the channel selection is done to a dataset from the healthy structure, in which PCA is applied. In that way, the most relevant and least redundant sensors are selected and its transformation is obtained.

In the second step, the channel selection is done to the incoming data just as in the section , but instead of using these data to detect damage, the data is transformed using PCA and the PCs extracted from the Feature Extraction part of the first step. After applying both of them, the damage detection is done.

Laboratory tower definition

The tested structure is a tower model, similar to those of a wind turbine. From Fig. 2a it can be seen the components of the structure: jacket, tower and nacelle. As a whole, this structure is 2.7m high. The tower is composed by three sections joined with bolts, while the jacket is composed with several sections, all of them are joined with bolts, with a torque of 12Nm. The top piece is 1m long and 0.6m

width. There, a modal shaker is placed to simulate the nacelle mass of the wind turbine. To simulate damages, a 5mm cracked section was placed in the jacket at different locations (see Fig. 2c), all these damages are defined in Table 1.

The modal shaker excites the structure using a white noise. The data is acquired using an OROS OR36 system. The vibration modes of the structure are calculated using Operational Modal Analysis. In this case, the first 10 modes are analyzed, which are located under 100 Hz.

Fig. 2: Data preprocessing methods for damage detection

Table 1: Damage Definitions

Damage	Location	Crack	Color in graph
D1.1	Floor 2 - Side 1	5mm	Red
D1.2	Floor 2 - Side 2	5mm	Dark-Blue
D2.1	Floor 1 - Side 1	5mm	Orange
D2.2	Floor 1 - Side 2	5mm	Pink
D2.3	Floor 1 - Side 3	5mm	Purple
D3.1	Floor 3 - Side 1	5mm	Brown
D3.2	Floor 3 - Side 2	5mm	Blue

Results

All the results are shown in Fig. 3. In each plot, each bar corresponds to a different time series which the height represents the damage indicator given by the NullSpace method. The first ten bars correspond to the healthy structure, the following five bars to the damage D1.1. After applying each type of damage, the healthy bar is placed again (next five bars) and the experiments are repeated. In this way, it is verified whether the damage indicator come back to the previous state. The different damages are shown in different colors (see Table 1 for more information). Besides, a horizontal line is shown in each plot, it represents the threshold, which determine if the experiment belongs to a healthy or damaged structure.

NullSpace Base Results. Fig. 3a shows the result of the NullSpace damage detection method by itself. The results are good, and all the damages are well detected. The only drawback is that after applying the damages D3.1 and D3.2 the structure is not returning to its initial state, and false positives appear in the results.

Feature Extraction Results. This is done in two different variants. Fig. 3b shows the results of applying PCA with no data reduction. It can be seen that the results are good, and all the damages are well detected, with no false positives. The other variant, the one shown in Fig. 3c, here the most significant Principal Components are chosen, those which retain the 99% of the cumulative variance. The results show that the damages in the upper part (D3.1 and D3.2) are not well detected after the reduction of dimensionality.

Feature Selection Results. In this case, a Feature Selection is done. Fig. 3d shows the result of selecting the eight most relevant sensors. Results are good, all the damage indicators are above the damage threshold, when a damage is present. However, sensitivity has decreased, and the damage indicator values are much smaller than the ones in previous solutions.

Mixed Method Results. Finally, the mixed case is shown in Fig. 3e. First the selection of variables is performed, the eight most relevant sensors are selected (the same sensors selected in the previous step) and later, PCA is applied to the selected features. This method gives good results as can be seen in the plot, by using half of the sensors used in the first solution, better results are achieved.

Conclusions and Future work

It is very known that the preprocessing of the original measurements from sensors in any structure is able to change the results of the damage detection algorithm. In this work, three different data preprocessing methods were introduced and applied before using a specific damage detection algorithm (NullSpace). Results of each method are analyzed and finally they are compared.

By applying these data preprocessing methods, a more sensitive and a more effective damage detection algorithm is obtained. The damage detection method has shown that is capable of detecting really small damages in the tested structure. Besides, if this method is supported by preprocessing methods, the results get better, and less information is used to build the model. Thanks to this reduction, the execution time gets smaller, and this helps to have a more effective damage detection method. Being able to detect damage using less sensors shows that the fact that having more sensors does not mean that the information they give is really used to detect damage. The best result is accomplished by applying Feature Selection and Extraction, specifically UmRMR for selecting sensors and PCA for extracting information.

In the future, these preprocessing methods will be applied to another damage detection algorithm. In that way, the validation of the Feature Selection/Extraction method will be guaranteed. On the other hand, and continuing with previous work by the authors [17], the preprocessing methods proposed here, will be applied in algorithms which include environmental and operational changes. Finally, the next challenge is to apply these methods to a real scale wind turbine.

Fig. 3: Data preprocessing methods for damage detection

References

- [1] Claus-Peter Fritzen. Vibration-Based Techniques for Structural Health Monitoring. In Structural Health Monitoring, volume 51, pages 45–224. ISTE, December 2006.
- [2] Ekhi Zugasti, Ana Gomez Gonzalez, Javier Anduaga, Miguel Angel Arregui, and Fernando Martínez. NullSpace and AutoRegressive damage detection: a comparative study. *Smart Materials and Structures*, 21(8):085010, August 2012.
- [3] Isabelle Guyon and André Elisseeff. An introduction to Feature Extraction. In Feature Extraction, volume 207 of Studies in Fuzziness and Soft Computing, pages 1–25. Springer, 2006.
- [4] I. T. Jolliffe. Principal Component Analysis. Springer Series in Statistics. Springer New York, New York, 2002.
- [5] a Hyvärinen and E Oja. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13(4-5):411–30, 2000.
- [6] John A. Lee and Michel Verleysen. Nonlinear Dimensionality Reduction. *International Statistical Review*, 76(2):308–309, August 2008.
- [7] Hervé Abdi. Partial least squares regression. *Encyclopedia of Measurement and Statistics*, 2007
- [8] M Dash and H Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.
- [9] Isabelle Guyon and A Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003
- [10] PR Cohen and David Jensen. Overfitting explained. *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.
- [11] Harald Stögbauer, Alexander Kraskov, Sergey a. Astakhov, and Peter Grassberger. Least Dependent Component Analysis Based on Mutual Information. *Physical Review E*, 70(Mic):18, May 2004.
- [12] F Fleuret. Fast binary feature selection with conditional mutual information. *The Journal of Machine Learning Research*, 5:1531–1555, 2004.
- [13] R Battiti. Using mutual information for selecting features in supervised neural net learning. *Neural Networks, IEEE Transactions*, 5(4):537–50, January 1994.
- [14] H Peng, F Long, and C Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence*, 2005.
- [15] Gavin Brown, A Pocock, MJ Zhao, and M Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 2012.
- [16] Junling Xu. Adapt the mRMR criterion for unsupervised feature selection. *Advanced Data Mining and Applications*, pages 111–121, 2010.
- [17] Ekhi Zugasti, Javier Anduaga, Miguel Angel Arregui, and Fernando Martínez. NullSpace Damage Detection Method with Different Environmental and Operational Conditions. *Proceedings of the 6th European Workshop of shm*, pages 1368–1375, 2012.