Gas Turbine Fault Diagnosis using Random Forests

Manolis Maragoudakis [1] and **Euripides Loukis** [1] and **Panayotis-Prodromos Pantelides** [1]

Abstract. In the present paper, Random Forests are used in a critical and at the same time non trivial problem concerning the diagnosis of Gas Turbine blading faults, portraying promising results. Random forests-based fault diagnosis is treated as a Pattern Recognition problem, based on measurements and feature selection. Two different types of inserting randomness to the trees are studied, based on different theoretical assumptions. The classifier is compared against other Machine Learning algorithms such as Neural Networks, Classification and Regression Trees, Naive Bayes and K-Nearest Neighbor. The performance of the prediction model reaches a level of 97% in terms of precision and recall, improving the existing state-of-the-art levels achieved by Neural Networks by a factor of 1.5%-2%.

1 INTRODUCTION

Development of effective Gas Turbine Condition Monitoring and Fault Diagnosis methods has been the target of considerable research in recent years. This is due to the high cost, sensitivity and importance of these engines for most industrial companies. Most of this research is directed towards the diagnosis of Gas Turbine blading faults, because of the catastrophic consequences that these faults can have, if they are not diagnosed in time. Even very small blading faults can very rapidly grow and result to huge destructions ([1], [2], [3]). Blading faults diagnosis is regarded to be a very difficult problem, because of the high levels of noise in all relevant measurements and the high interaction between the numerous Gas Turbine blading rows. Therefore, it is very important to take advantage of the processing power of modern computers, in order to provide a fast and reliable engine condition diagnosis from available measurements and to develop the highest possible level of intelligence and assistance to the operation and maintenance personnel. The Gas Turbine Blading Fault Diagnosis problem was originally addressed in [4] and [5], based on classical pattern recognition methods. Our contribution to the domain, is the introduction of an ensemble classifier, namely Random Forests, for the first time for the task at hand, which outperforms all previous attempts to Gas Turbine Blading Fault Diagnosis. Furthermore, Random Forests can provide some insight on the interrelationships between input features, unlike Neural nets, thus directing domain experts at selecting which measurement tools to use in real world applications.

2 PROBLEM & DATA DESCRIPTION

The present work is based on data acquired from dynamic measurements on an industrial Gas Turbine into which different faults were artificially introduced. During the experimental phase four categories of measurements were performed simultaneously:

1. Unsteady internal wall pressure (using fast response transducers P2 to P5).
2. Casing vibration (using accelerometers A1 to A6 mounted to the outside compressor casing).
3. Shaft displacement at compressor bearings (using transducer B).
4. Sound pressure levels (using double-layer microphone M).

Five experiments were performed, testing the datum healthy engine and a similar engine with the following four typical small (but quite rapidly growing, as mentioned in the introductory section) and also not straightforwardly diagnosable faults:

1. Fault-1: Rotor fouling.
2. Fault-2: Individual rotor blade fouling.
3. Fault-3: Individual rotor blade twisted (by appr. 8 degs).
4. Fault-4: Stator blade restaggering.

Tests were performed at four different engine loads (full load, half load, quarter load and no load), both for the healthy engine as well as for the above four faults. At each load, four series of time-domain data were acquired for each instrument (two series in each of the two sampling frequencies, l = 13 kHz and m = 32 kHz). 12 different measuring instruments were used and measurements were taken for every possible combination between engine's 5 operational conditions (healthy engine and 4 faulty conditions), 4 different engine loads (full load, half load, quarter load and no load) and 2 sampling frequencies (low and high). To be more precise, regarding engine's healthy condition, measurements have been taken for every combination between the engine load and sampling frequency (total 8 different combinations). Especially in engine's faulty condition there's been one more measurement series for all the above combinations. Consequently, for every instrument we have aggregately 72 different measurements: 8 healthy engine's measurements and 64 faulty engine's measurements. For every instrument, each and every one of the above measurements consists of 27 values that are forms of the spectral difference of the first 27 harmonics of rotor's shaft rotational frequency. So, if we would like to present the entirety of data in a data base then this would be composed of 864 instances described by 27 distinct attributes, corresponding to the 27 harmonics.

3 RANDOM FORESTS

Despite the fact that Random Forests have been quite successful in classification and regression tasks, to the best of our knowledge, there has been no research in using the afore-mentioned algorithm for Gas

[1] University of the Aegean, Department of Information and Communication Systems Engineering, Samos, Greece

Turbine Fault Diagnosis. Random Forests are a combination of tree classifiers such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. A Random Forest multi-way classifier Θ(x) consists of a number of trees, with each tree grown using some form of randomization, where x is an input instance [8]. The leaf nodes of each tree are labeled by estimates of the posterior distribution over the data class labels. Each internal node contains a test that best splits the space of data to be classified. A new, unseen instance is classified by sending it down every tree and aggregating the reached leaf distributions. In order to make the classification process more formal, suppose that the joint classifier Θ(x) contains x individual classifiers Θ_1(x), Θ_2(x),...,Θ_x(x). Let us also assume that each data instance is a pair *(x,y)*, where x denotes the input attributes, taken from a set A_i, i=1,...,M and y symbolizes the set of class labels L_j, j=1,...,c (c is the number of class values). For reasons of simplicity, the correct class will be denoted as y, without any indices. Each discrete attribute A_i takes values from a set V_i, i=1 to m_i (m_i is the number of values attribute A_i has). Finally, the probability that an attribute A_i has value v_k is denoted by $p(v_{i,k})$, the probability of a class value y_j is denoted by $p(y_j)$ and the probability of an instance with attribute A_i having value v_k and class label y_j is symbolized by $p(y_j|v_{i,k})$.

Each training example is picked up from a set of *N* instances at random with replacement. By this procedure, called bootstrap replication, a pool of 36.8% of the training examples are not used for the tree construction phase. These out-of-bag (oob) instances allow for computing the degree of strength and correlation of the forest structure. Suppose that O_k(x) is the set of *oob* instances of classifier Θ_k(x). Furthermore, let $Q(x, y_j)$ denote the subset of *oob* samples which were voted to have class y_j at input example x. An estimate of $p(\Theta(x) = y_j)$ is given by the following equation:

$$Q(x, y_j) = \frac{\sum_{k=1}^{K} I(\Theta_k(x) = y_j; (x, y) \in O_k)}{\sum_{k=1}^{K} I(\Theta_k(x); (x, y) \in O_k)} \quad (1)$$

where I(·) is the indicator function.

The margin function which measures the extent to which the average vote for the right class y exceeds the average vote for any other class labels is computed by:

$$margin(x, y) = P(\Theta(x) = y) - max^{c}_{j=1, j\neq y}(P(\Theta(x) = y_j) \quad (2)$$

Since strength is defined as the expected margin, it is computed as the average over the training set:

$$s = \frac{1}{n} \sum_{i=1}^{n} (Q(x_i, y) - max^{c}_{j=1, j\neq y} Q(x_i, y_j)) \quad (3)$$

The average correlation is given by the variance of the margin over the square of the standard deviation of the forest:

$$\overline{p} = \frac{Var(margin)}{\sigma(\Theta())^2} \quad (4)$$

is estimated for every input example x in the training set $Q(x, y_j)$.

4 EXPERIMENTAL RESULTS

We applied two versions of Random Forests (Random Input (RI) Forests and Random Combination (RC) Forests) on the Gas Turbine data set, using oob estimates. As for evaluation metric, we considered per class precision and recall. Accuracy in some domains, such as the one at hand, is not actually a good metric due to the fact that a classifier may achieve high accuracy by simply always predicting the non faulty class. This problem particularly appears in the present task, where, from more than 2/5 of the data set contained the aforementioned class. A set of well-known machine learning techniques have constituted the benchmark to which our results have been compared: Multi-layer Perceptron Neural Networks, Naive Bayes, Classification and Regression Trees (CART), and k-Nearest Neighbor (k-NN) instance-based learning. Cross validation was performed with k-NN in order to determine the best k. As regards to the Random Forests implementation, the best results were obtained by using 500 trees and 6 features. Due to lack of space, the evaluation outcome is depicted in the following figure, for the precision metric (F1 to F4 denotes the fault categories and OK denotes the non faulty state).

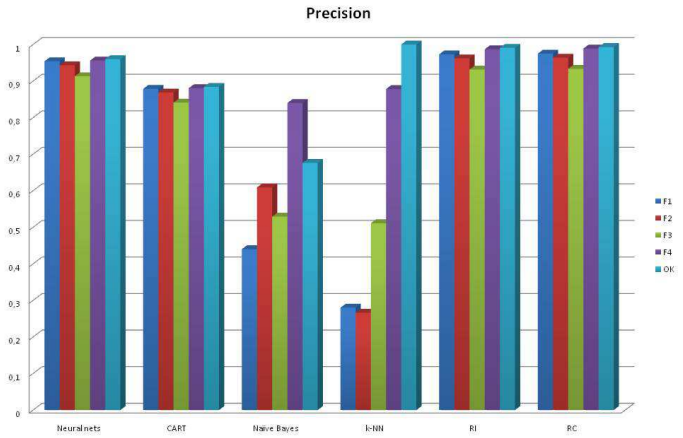

Figure 1. Evaluation results in terms of precision for all methodologies.

REFERENCES

[1] E. Loukis, P. Wetta, K. Mathioudakis, A. Papathana siou, K. Papailiou, Combination of Different Unsteady Quantity Measurements for Gas Turbine Blade Fault Diagnosis, 36th ASME International Gas Turbine and Aeroengine Congress, Orlando, 1991, ASME paper 91- GT-201.

[2] E. Loukis, Contribution to Gas Turbine Fault Diagnosis Using Methods of Fast Response Measurement Analysis, Doctoral Thesis, Athens, National Technical University of Athens, 1993.

[3] G. Merrington, O. K. Kwon, G. Godwin, B. Carlsson, Fault Detection and Diagnosis in Gas Turbines, ASME Journal of Engineering for Gas Turbines and Power, 113, 1991, 11-19.

[4] E. Loukis, K. Mathioudakis, K. Papailiou, A procedure for Automated Gas Turbine Blade Fault Identification Based on Spectral Pattern Analysis, Journal of Engineering for Gas Turbines and Power, 114, 1992, 201-208.

[5] E. Loukis, K. Mathioudakis, K. Papailiou, Optimizing Automated Gas Turbine Fault Detection Using Statistical Pattern Recognition, Journal of Engineering for Gas Turbines and Power, 116, 1994, 165-171.

[6] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Classification and regression trees. Wadsworth Inc., Belmont, California, 1984.

[7] Leo Breiman. Bagging predictors. Machine Learning Journal, 26(2):123140, 1996.

[8] Igor Kononenko. Estimating attributes: analysis and extensions of Relief. In Luc De Raedt and Francesco Bergadano, editors, Machine Learning: ECML-94, pp. 171182. Springer Verlag, Berlin, 1994.