Electrical cables diagnostics using an experimental dataset of Partial Discharge measurements containing contradictory patterns

Piero Baraldi, M. Compare, Enrico Zio, M. De Nigris, G. Rizzi

Electrical cables diagnostics using an experimental dataset of Partial Discharge measurements containing contradictory patterns

P. Baraldi[1], M. Compare[1], E. Zio[1,2]
[1]*Energy Department, Politecnico di Milano, Milan, Italy*
[2]*Ecole Centrale Paris-Supelec, Paris, France*

M. de Nigris, G. Rizzi
Enea Ricerca sul Sistema Elettrico – ERSE, Milan, Italy

ABSTRACT: Partial Discharge (PD) measurements have been proposed as a relatively economic and simple-to-apply experimental technique for retrieving information on the health state of an electrical cable. A set of PD measurements have been collected by Enea Ricerca sul Sistema Elettrico (ERSE), for building a diagnostic system of electrical cable health state. These experimental data may contain contradictory information which remarkably reduce the performance of the state classifier. In the present work, a novel technique based on the Adaboost algorithm is proposed for identifying contradictory PD patterns within an a priori analysis aimed at improving the diagnostic performance. Adaboost is a bootstrap-inspired, ensemble-based algorithm which has been effectively used for addressing classification problems.

1 INTRODUCTION

The state of components such as electrical cables may be difficult to diagnose unless destructive or very expensive tests are used. To overcome this, Partial Discharge (PD) measurements are considered as indicators of localized defects in electrical cables. During past experimental campaigns, Enea Ricerca sul Sistema Elettrico (ERSE) has built a database containing the values of the PD measurements and the corresponding health state of the cable. This database contains thousands of PD patterns recorded by a software tool that processes the PD measurements when these are performed and classified by experts on the basis of both their experience and ERSE guidelines; for a small number (43) of them, the classification of the degradation state is guaranteed based on visual inspection. This is, in fact, a very expensive task which can be performed only occasionally since it entails the extraction of the cable from the ground and it implies the unavailability of the corresponding electrical line for several hours.

Based on these observed data, an empirical classifier can be developed for relating the PD measurements (input) with the health state of the cable (output, categorized into two classes: 'Bad' and 'Good'). On the other hand, errors can occur in data collection, i.e., when acquiring the measurements (e.g., bad sensors, operator errors, data transferring errors etc.), when processing the data (e.g., transcribing, transmission, omission errors, etc. (U.S. Statistical Policy Office, 2001)), when handling the databases, etc. Any erroneous data could undermine the informational content of a database, and recorded incoherent patterns could bias the mapping function created by training the classification algorithm, thus affecting its performance.

Incoherence in the present work is considered due to two main sources:

1) in some cases, information relevant to cable classification is missing. This leads to the fact that, without knowledge of the missing information, it is possible that two input patterns with the same values can be associated to both 'Good' and 'Bad' cables in the dataset. For example, the information regarding the insulation material of the tested cables has not been reported in the database, although it is relevant for cable classification since it influences the sensitivity of the cable to the discharges. Figure 1 gives a sketch of this type of incoherence. For the sake of simplicity, a mono-dimensional in-

put space is considered, that is, a generic parameter representative of the PD measurement is reported in abscissa in arbitrary units. For representation purposes, it is also assumed that cables whose insulation material is oil-paper are characterized by a value of the parameter smaller than a threshold (square marker in Figure 1) in case they are of class 'Good' (crosses), and larger in case they are of class 'Bad' (circles). If all the empirical patterns available were from oil-paper cables, it would be possible to build a classifier specific for this family of cables. However, ERSE experts believe that there are few patterns in the database that refer to cables whose insulation material is Cross Linked Poly-Ethylene (XLPE) or Ethylene Propylene Rubber (EPR). These families of cables are characterized by different relationships between the input parameters and the class; in the example of Figure 1, this situation is represented by assuming that the threshold that distinguishes between "Good" and "Bad" cables has a lower value than in the case of oil-paper cables. Notice that if the information regarding the insulation material is not reported, then the PD patterns of XLPE and EPR cables are assumed to be 'oil-paper' and thus may result as incoherent: Figure 1 shows that this situation of missing information results in the projection of patterns at different ordinates (missing information) on the ordinate of the oil-paper with the consequent introduction of some incoherent patterns.

2) there are degradation mechanisms that are not reflected in the PD measurements, i.e., some cable defects cannot be diagnosed using PD measurements, but require other investigation techniques. This results in the presence in the database of cables classified as 'Bad' on the basis of visual inspection, although the obtained PD measurements do not reveal any local defect and thus are typical of "Good" cables.

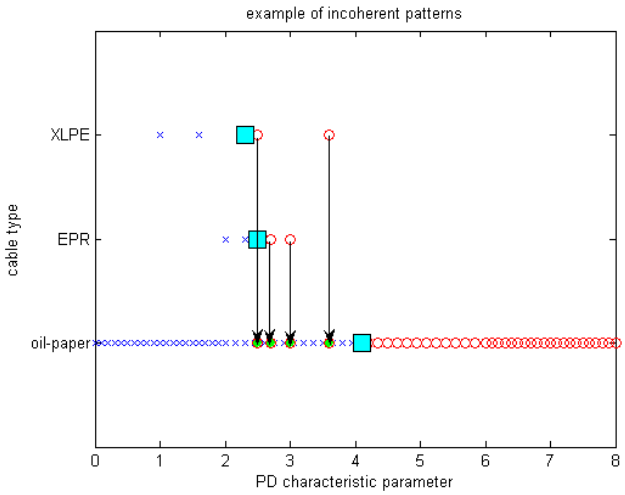

Figure 1: example of source of incoherent patterns.

Both situations described above relate to missing information in the database, i.e., variables related to the PD measurements which are available at the moment of the test but not recorded (case 1) or values of signals different from those measured in the experimental tests (case 2). The lack of such information renders some patterns of the dataset incoherent and contradictory, in the sense that identical or very similar values of the input signals are associated to different classes. The contradictory patterns in the dataset are obviously harmful for the development of empirical models that are trained on the basis of input-output data.

The objective of the present work is to propose a novel methodology (based on the Adaboost technique (Freund & Shapire 1997)) which allows identifying the contradictory patterns in a dataset, so that they can be eliminated or corrected before use for training of the classifier. In the example of Figure 1, the EPR and XLPE contradictory patterns would be identified and removed from the dataset, thereby allowing the development of a classifier of oil-paper cables.

It seems worth emphasizing that the problem here tackled differs from that of developing and using empirical classifiers in case of datasets containing missing or corrupted values, which has been largely discussed in the literature (Ho 1998, DePasquale & Polikar 2007, Polikar 2007). Approaches to these "missing feature/missing data problems" exist, which involve estimating the values of the missing data by exploiting the presence of some patterns complete with measurements of the signals missing in other patterns. On the contrary, in the problem of

interest in this work, measurements of the missing signals are not available in any pattern of the dataset, so that it is not possible to empirically infer them from available complete patterns. Thus, the solution inevitably adopted entails the removal from the dataset of the patterns that cannot be univocally classified, due to the missing information. This cleaning of the dataset is expected to improve the performance of the algorithm in the diagnosis of the health state of those cables which can be coherently classified on the basis of the available information.

The paper is organized as follows: in Section 2, a brief outline on the characteristics of PD measurements is provided; in Sections 3 and 4, the key ideas underlying the considered classification algorithms are explained; in Section 5, the methodology is applied on the available PD measurement dataset. Finally, Section 6 concludes the paper with some considerations.

2 PD MEASUREMENT

The PD measurement is acquired through an off-line process. First, the cable under inspection is de-energized and disconnected from any source or load from all terminals. The PD measurement system, equipped with its own generator, is then connected to energize the cable with damped oscillatory voltages of frequency in the range 150 to 250Hz. In terms of signal acquisition, the system captures the high frequency signal generated by the partial discharges activity and conveys it to a signal conditioner unit through the coupling capacitor. The signal conditioner reduces the overall bandwidth of the acquired wave and amplifies it, thus enhancing its signal-to-noise ratio.

Schematically, the PD measurement is performed in three successive phases corresponding to three energizing voltage levels; in the first phase, the PD inception voltage U_i (i.e., the voltage value at which the PDs start) is reached through a stepwise or continuous increase of the voltage applied to the cable; at this voltage level, the following two parameters are collected:

1. The PD value, i.e. the value of the discharge expressed in pC.
2. The dispersion index, i.e. the length of the cable in which the discharge activity is localized.

In the following two phases, the cable is tested at the nominal voltage level U_0 and the maximum voltage value U_{max}, respectively, and the values of the above two parameters are measured. For convenience of data manipulation, the values of U_i, U_0 and U_{max} are normalized with respect to U_0.

The three triplets of values (normalized voltage level, PD value and dispersion index) corresponding to the three different values of voltage, constitute a pattern of 9 features in which the normalized nominal voltage value U_0/U_0 is always equal to 1; thus, this feature is non-discriminating and for this reason it is not considered in the diagnostic analysis. A PD measurement pattern is then made up of 8 feature.

During past campaigns, a set of 43 PD measurement patterns has been collected by ERSE. The dataset contains 16 patterns of class ‘Bad’ and 27 patterns of class ‘Good’. The classification of the health state (diagnosis) is based on visual inspections made by ERSE experts who extracted from the ground and cut up a cable section after acquiring the PD measurement patterns. For these 43 patterns, paper reports have been prepared by the experts, containing photos and further information about the tested cables and the electrical line which they belong to.

3 CLASSIFIER

A diagnostic system based on the ERSE dataset has been built by using the Evolutionary Fuzzy C Means (EFCM) as classification algorithm (Yuan & Klir 1997). This is a supervised classification algorithm that uses the knowledge of the class of the patterns for finding for each class an optimal Mahalanobis metric that defines a geometric cluster as close as possible to the a priori known class. The Mahalanobis metrics are defined by the matrices whose elements are identified by the supervised evolutionary algorithm so as to minimize the distances between the patterns of each class and the center (also referred to as cluster prototype) of the corresponding cluster. Further details on the EFCM algorithm can be found in (Yuan & Klir 1997) and (Zio & Baraldi 2005).

4 IDENTIFICATION OF CONTRADICTORY PATTERNS

Let x_k, k=1,…, n, be a pattern of an empirical dataset S. The information available for each pattern are the values of the f features and its class c_k, i.e. $x_k=(x_{1k}, x_{2k}, ..., x_{fk}, c_k)$, c_k=1,2, …, Ω (in the present case study Ω=2). Let us assume that in the dataset S there is an unknown number nc of contradictory patterns, i.e. patterns of different classes with very similar input values.

The methodology here proposed for the identification of the contradictory patterns is based on the assumption that for an empirical classification algorithm it is difficult to learn the relationships between the input signals and the class of the patterns in those zones of the input space characterized by the presence of contradictory patterns.

Different types of classification algorithms give different warnings of their difficulties in learning the training dataset S. This Section investigates the behavior of the Adaboost classification approach in the case in which a dataset S containing contradictory patterns is used to train the classifiers.

4.1 *Adaboost Algorithm*

The main characteristics of the Adaboost algorithm are suggested by its name which stands for Adaptive Boosting. It is a boosting algorithm: a sequence of B classifiers, C_b, b=1,2,…B, is created by training a classifier algorithm on different bootstrap samples S_b*, b=1,2,…, B. The probability mass distribution, $D_b=\{p_b(1), p_b(2),…, p_b(n)\}$, whose generic element $p_b(k)$ gives the probability of drawing pattern x_k from S in the bootstrap sample S_b*, is opportunely altered after building a classifier in order to ensure that more informative points are drawn into the next dataset used for building the successive classifier. In this sense, Adaboost is adaptive because it updates the distribution D such that after a classifier C_b is built, the subsequent classifier, C_{b+1}, pays more attention to training patterns that were misclassified by C_b. In particular, if pattern x_k is misclassified by the generic classifier C_b, then the probability p_{b+1} (k) that x_k is drawn when building the next data training set (S_{b+1}*) is increased with respect to $p_b(k)$; on the contrary, sampling probabilities of the points correctly classified are reduced. In this way, the probability that S_{b+1}* will contain a larger number of patterns x_k increases and this gives the classifier C_{b+1} more chances to correctly classify x_k. In case x_k is again misclassified, then $p_{b+2}(k)$ will be further enhanced. The increasing behavior of the sampling probability associated to x_k ends when a classifier that is able to correctly classify x_k is built. In this way, subsequent classifiers are tweaked in favor of those patterns misclassified by previous classifiers and thus tend to have higher performance on these difficult patterns.

The classifiers are then combined through weight majority voting to obtain the final classification. The voting weight of a classifier is strictly dependent on its performance: the larger the number of patterns of S correctly classified the larger its vote.

4.2 *Degree of contradictoriness*

Within an Adaboost classification approach, contradictory patterns are expected to be among the patterns that are misclassified by the ensemble classifiers and thus with an associated high value of the probability mass functions $p_b(k)$, b=1, 2,…, B. The idea is thus to consider as indicator of the degree of contradictoriness of pattern x_k, k=1, 2,…, n, the quantity:

$$w_k = \frac{\sum_{b=1}^{B} p_b(k)}{B} \quad (1)$$

Given the updating dynamics of the distribution D, a pattern x_k which is correctly classified by all classifiers C_b, b=1,…,B, is associated to low values of $p_b(k)$, b=1,…,B. On the contrary, the probability masses $p_b(k)$ associated to patterns which are difficult to be classified have the oscillating behavior described in Section 3.1; thus, the mean value w_k of the probability masses associated to these patterns tends to be larger: the contradictory patterns of S are then expected to occupy the first positions of the ranking of the values $w_1, w_2,…,w_n$.

Once the mean values w_k, k=1,2,..,n, of Equation 1 have been computed, the following criterion can be applied for identifying a set $\hat{S}_c$ of nr patterns candidates to be contradictory and thus removed from the dataset: compute the mean W and the standard deviation V of the vector [w_1, w_2,…, w_n] and then

consider as candidates to be contradictory the patterns (if any) of the set $\{k \,|w_k \geq W+V\}$.

5 APPLICATION TO THE ERSE PD MEASUREMENT DATASET

The present Section reports the results obtained by applying the methodology proposed above for the classification of contradictory patterns to the ERSE PD measurement dataset described in Section 2.
The proposed methodology identifies 7 out of the 43 patterns of the dataset as contradictory. Once the 7 selected patterns have been removed from the dataset, the Adaboost algorithm can be adopted to build the final diagnostic system trained on the set S' of the remaining 36 patterns. The results obtained have been compared with those obtained by a diagnostic system trained with the set *S* of all the available 43 patterns (worst case).

In order to get a reliable estimation of the classification performances, the Leave-One-Out (LOO) cross validation scheme has been applied.

In the LOO approach, an instance is omitted from the training sample; when the classifier is built, the prediction (correct or incorrect) for the omitted instances is obtained; the process is repeated for all the instances in the training sample; the estimation of the true error is given by the proportion of instances incorrectly classified. This estimator has low bias but its variance tends to be large.

Table 1 summarizes the obtained performance: the removal of the patterns identified as contradictory by the proposed methodology results in a remarkable increase of the performance of the diagnostic system, which in the present case becomes infallible.

Table 1: performance of the diagnostic system.

Training Set	Performances on Test Set S' (36 patterns)
S (43 patterns)	0.611
S' (36 patterns)	1

6 CONCLUSIONS

Errors in data collection can result in databases containing contradictory patterns which could bias the mapping function created by training a classification algorithm; this can significantly affect the classification performance.

An original methodology which allows to recognize contradictory patterns has here been proposed and applied with satisfactory results to the PD measurements dataset collected by ERSE for diagnosing the health state of electrical cables.

REFERENCES

R. Polikar (2007): *Bootstrap Inspired Techniques in Computational Intelligence.* IEEE Signal Processing Magazine. Vol. 24 No. 4, pp. 59-73

B. Efron (1979): *'Bootstrap Methods: Another look at the Jackknife'.* Annals of Statistics. Vol.7, no. 1, pp. 1-26.

R.E. Shapire (1990): *The Strength of weak learnability. Machine Learning*. Vol. 5 no 2, pp. 197-227.

Y Freund, R.E. Shapire (1997): *'A Decision-Theoretic generalization of On line Learning and Application to Boosting'*. Journal of computing and system Science. Vol. 55, pp. 119-139.

E. Zio (2009): *'Computational methods for reliability and risk analysis'*. World Scientifics.

P.J. Boland (1989*): 'Majority Systems and the Condorcet jury problem'*. Statistician. Vol. 30, no. 3, pp. 181-189.

B. Yuan, G. Klir (1997): *'Data driven identification of key variables'*, in D. Ruan (Ed.), Intelligent Hybrid Systems Fuzzy Logic, Neural Network, and Genetic Algorithms, Kluver Academic Publishers, pp. 161–187 (Chapter 7).

E. Zio, P. Baraldi (2005): *'Identification of Nuclear Transients via Optimized Fuzzy Clustering'*, Annals of Nuclear Energy, Vol. 32, No. 10, pp. 1068-1080.

T.K. Ho (1998): *'Random Subspace method for constructing decision forests'*. IEEE Trans. Pattern Anal. Machine Intell. Vol. 20, No. 8, pp. 832-844.

J. DePasquale, R. Polikar (2007): Random Feature Subset Selection for Ensamble Based Classification of Data with Missing Features. Lecture Notes in Computer Science, Vol 4472. M. Haindl and F. Roli, Eds Berlin: Springer-Verlag. pp. 251-260

Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget (2001): *'Statistical Policy-Working Paper 31: Measuring and Reporting Sources of Error in Surveys'.*