

Classification of Voltage Sags based on k-NN in the Principal Component Space

J. Meléndez¹, X. Berjaga¹, S. Herraiz¹, J. Sánchez² and M. Castro²

¹ Institut d'Informàtica i Aplicacions
eXiT., Universitat de Girona

Campus Montilivi, 17071 Girona (Spain)

Phone/Fax number:+0034 972 418391, e-mail: quimmel@eia.udg.edu, xberjaga@eia.udg.edu, sherraiz@eia.udg.edu

² Power Quality Department of Endesa Distribución, Barcelona (Spain)

Mailing address

Phone, fax, jslosada@fecsa.es, MCastro@enher.es

1. Brief introduction

A new method for the classification of sags is proposed. The goal is to deduce the origin of sags (upstream or downstream of the transformer) registered in distribution substations. The method is based on the existence of a case base of previous registers which origin is well known. This case base is used to infer the origin of new sags based on the retrieval of similar sags using a distance criterion. This distance computed in the principal component space is also used in the decision step to decide the origin of the new sag.

Key words: Fault location, Principal Component Analysis (PCA), Classification

2. The proposed methodology

We introduce a new method to assist power quality monitoring for classification of electrical sags gathered in distribution substation in order to determine their origin. PCA is proposed to model datasets of sags which origin, in MV or HV, are known. A PCA model built with existing sags originated in HV is built and later used to identify new sags by projecting them against this model and evaluating similarity of new sags in this new space. The goal is to capture the relevant information of sags originated in HV useful to discriminate them from those originated in MV. Next we present the ways for preparing the data as well as the whole procedure in details. The procedure steps could be broken down in two general stages: (i) case base preparing and construction in the principal component space, and (ii) model exploitation. The tasks to perform the first stage include the creation of the PCA model and the projection of the existing registers (voltage and current waveforms) into the principal component space. This is performed with a training set of data which origin is known. Here, they are described in more details following the execution order:

- *RMS Value Computations:* Instantaneous RMS value for each variable (three voltages and three

currents) is computed.

- *Autoscaling:* Since RMS magnitudes of voltages and currents are completely different autoscaling is a preprocessing stage needed to avoid overweighting of voltage variables towards currents. It results in zero-mean centered data with a unit variance.
- *Model Creation:* HV-PCA model is created using training subsets.
- *Projection:* HV and MV sags are projected into HV-MPCA model resulting in a case base in the principal components space.

The case base obtained in the new space presents an important reduction on the number of variables that will be taken into account. Moreover, this new variables are independent, which means that the Euclidean distance will reflect the differences between variables considering the correlation among them. At the same time, the use of a projection operator that models one class of data produce a better separation of cases in the new space, also allowing the use of well known statistics as Q and T² in this space to evaluate similarity between new sags and those in the case base.

The second stage, the exploitation of this case base built in the projection space, is based on the similarity criteria between new sags and those in the case base previously diagnosed. Thus, the methodology to classify sags follows these steps:

- *Projection of new sags:* The same HV-PCA model is used to project new sags in the principal component space.
- *Retrieval of the k-NN:* A subset of k nearest neighbors is identified based on two criteria: Q and T² statistics. Firstly, a larger selection of candidates (k1) is made based on its adequacy to the model using the Q statistic. From it, the k2 best candidates are selected by comparing the T² statistic (distance to the center of the model).
- *Class Determination:* The class of the sag is determined by the comparison between a threshold and the relation of the distances of all the retrieved cases that belong to the model (its class is HV) and

all the distances that have been obtained.

Results obtained with this methodology have been validate using 5-fold cross validation and performance measured with ROC (Receiver Operation Characteristics) curves.

3. Classification of Sags gathered in distribution substations

The methodology has been tested using different pairs of values for the k_1 and k_2 retrieved cases and using different thresholds to compute the ROC curve depicted in Fig. 1. It can be observed that all the tests had a very good performance: AUC near 1. The red dashed line represents a random classifier. According to AUC (Table 1) criteria, the best classifier is the one using $k_1 = 20$ and $k_2 = 3$ because it has the greatest value, also having the nearest operating point to (0,1) in the curve.

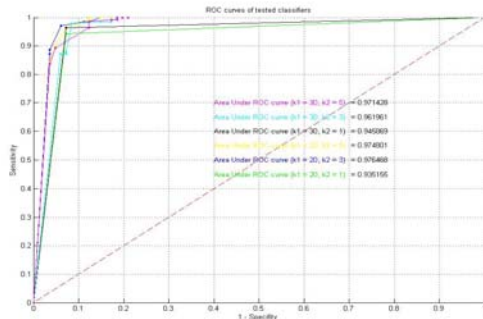


Fig. 1. ROC curves and AUC of several classifiers

Table 1: AUC values of the classifiers

(k1, k2)	AUC
(20,5)	0.975
(20,3)	0.976
(20,1)	0.935
(30,5)	0.971
(30,3)	0.962
(30,1)	0.946

Once the best parameter for the classifier is known, the next step is to choose the best decision threshold. As exposed before, the easiest way is to find the nearest point to the point (0,1) because there are no errors in the classification. For this classifier the Sensitivity and specificity of Table 3 have been obtained.

Table 2: Table of Sensitivities and 1 - Specificities of the best tested classifier

Threshold	1 - Specificity	Sensitivity
1	0	0
0,8	0,036	0,872
0,9	0,036	0,872
0,7	0,036	0,886
0,6	0,061	0,971
0,5	0,111	0,986
0,4	0,124	0,986
0	0,148	1
0,1	0,148	1
0,2	0,148	1
0,3	0,148	1

As shown in Table 2, the classifier has a better classification ratio with low threshold values (threshold with a maximum value of 0.3). So, any of these values can be chosen as the best operating point to classify sags according to HV and MV.

6. Conclusions

The methodology presented in this paper for sag classification has been demonstrated to be very good with the data set used in the experiments. The usage of the ROC curves and the AUC has been used to compare different configurations of the classifier with excellent performance and have been used to decide the best configuration for this location of origin of sags.

Although these good results encourages to proceed improving the methodology, the data used only represents a part of all the available cases, and future experiments will be done in order to evaluate the correlation among the subset used in this paper and the whole set of cases. Another point to take into account will be to use more distances in the k-NN selection step.

Acknowledgement

This research has been made possible by the interest and participation of ENDESA DISTRIBUCION and its Power Quality Department. It has also been supported by the research projects DPI2005-08922-C02-02 and DPI2006-09370 funded by the Spanish Government.

References

- [1] Mora, J.; Llanos, D.; Melendez, J.; Colomer, J.; Sanchez, J. and Corbella, X., "Classification of Sags Measured in a Distribution Substation based on Qualitative and Temporal Descriptors", in 17th Int. Conf. On Electricity Distribution, May 12-15, Barcelona, Spain, 2003.
- [2] De Mantaras R. L. and Plaza E., "Case-based reasoning: An overview", in AI Communications, Vol. 10, No 1, pp 21-29.
- [3] Aamodt A., Plaze E., "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches", in AI Communications 1994, Vol. 7, No. 1, pp. 39-59.
- [4] Kohavi R., "A study of cross-validation and bootstrap for accuracy estimation and model selection", in Proceedings of 14th International Joint Conference on Artificial Intelligence, Vol. 2, No. 12, pp. 1137-1143.
- [5] Bradley A. P., "The use of the area under the ROC curve in the evaluation of machine learning algorithms", in Pattern Recognition, Vol. 30, No. 7, pp. 1145-1159.
- [6] Therrien C. W., Decision Estimation and Classification: An introduction to pattern recognition and related topics, Wiley, 1989.
- [7] Fawcett T., Pattern Recognition, Letters, 2006, pp. 861-874.