Comparison Between Random Forest Algorithm and J48 Decision Trees Applied to the Classification of Power Quality Disturbances

Fábbio A. S. Borges, Ricardo A. S. Fernandes, Member, IEEE, Lucas A. M. and Ivan N. Silva, Member, IEEE

***Abstract*— This paper presents a methodology for the classification of disorders related to the area of Power Quality. Therefore, we used the Random Forest algorithm, which corresponds to an effective data mining technique, especially when dealing with large amounts of data. This algorithm uses a set of classifiers based on decision trees. In this sense, the performance of the proposed methodology was evaluated in a comparative way between the Random Forest and the type J48 Decision Tree. For this analysis to be possible, synthetic electrical signals were generated, where this disturbances were modeled through parametric equations. After the performance analysis, it was observed that the results were promising, since the Random Forest algorithm has the best performance.**

Index Terms— Random Forest, power quality disturbances.

I. Introduction

Disturbances related to the area of Power Quality (PQ) are characterized by changing the waveforms of sinusoidal voltage and current, which can affect the operation of certain equipment [1]. Among these disturbances, there are sags, swells, interruption, harmonic distortion, oscillatory transients. Such disturbances are becoming a problem for both the power utilities as well as consumers, making it necessary to eliminate or mitigate the cause of their occurrence in order to ensure good power quality.

Thus, the detection and classification of disturbances is a primary task so that measures to control and mitigate disturbances can be adopted. However, this is no easy task, because the identification of these disturbances often require the analysis of a large amount of data measured by equipment installed on the network, besides the fact that many of the disturbances have similar features.

In this context, it is desirable to employ data mining tools, because they can identify these PQ disturbances in a fast and automated manner. Additionally, it is desirable that these tools are able to analyze a large volume of data and to acknowledge a pattern in the data in order to relate it to a possible disturbance.

The area of disturbance detection and classification has been the subject of several studies in recent years [2]. These studies utilize techniques to extract relevant signal characteristics, they reduce the dimensionality of the input data and remove redundant features of the original vector. The extracted features are then used as inputs to a method of pattern classification responsible for relating an input vector with a disturbance. Among the most used methods we highlight Fuzzy Logic, Artificial Neural Network and Support Vector Machine (SVM).

Following the above context, this work proposes the Random Forest algorithm with the interest to hold a review/classification for a database composed of waveforms that contain power quality disturbance. Random Forest is developed by Leo Breiman [3]. RF fits many classification trees to a data set and then combines the prediction from all the correlated trees. Each tree depends on the value of a separately sampled random vector.

During the feature extraction step, time calculations on time domain that have low computational effort are used. Following, the feature vector is used as the RF input so that the final response is defined by the class that has the highest number of outputs, that is, by the account of the outputs presented by each of the decision trees that compose the algorithm. Finally, the classification results are obtained and compared with the response of a Decision Tree (DT) that uses the training algorithm J48.

II. DATABASE COMPOSED OF SYNTHETIC SIGNALS

The objective of the database modeling is to store the maximum number of signals with different characteristics of the disturbance. These signs will be used to test the proposed methodology. In this work, the occurrence of the following disturbances was considered: voltage sags, swells, flickers, harmonic distortion, voltage interruptions, oscillatory transients, voltage sags in conjunction with harmonic distortion and swells together with harmonic distortions. A database was created consisting of windows obtained for synthetically modeled disturbances, based on mathematical models proposed in [4].

Therefore, the windows that make up the database were derived from 100 case studies for each disturbance, and each of the disturbances has a total of 10 cycles at nominal frequency of 60 Hz and sampled rate of 128 points per cycle. This windowing occurs through the shifting of the data window (which is the size of one cycle of the signal) in steps of 1 point until it covers the entire length of the signal.

Fábbio A. S. Borges, Lucas A. M and Ivan N. Silva are with the Department of Electrical and Computing Engineering, University of São Paulo, São Carlos, 13566-590, Brazil (e-mail: fabbioanderson@gmail.com, lucas.moraes@usp.br, insilva@sc.usp.br). Ricardo A. S. Fernandes is with the Department of Electrical Engineering , Federal University of São Carlos, 13565-905, São Carlos, Brazil (e-mail: ricardo.asf@ufscar.br)

The result of the process is the construction of a database comprised of approximately: 14864 sag windows, 12671 swell windows, 19706 flicker windows, 34084 harmonic distortion windows, 13277 harmonic distortion with windows, 12769 harmonic distortion with swell windows, 10366 interruption windows and 5836 transient windows

III. FEATURE EXTRACTION FROM THE WINDOWED SIGNALS

As soon as a disturbance is detected, the classification step is activated, which uses a stage of extraction of features combined to a decision tree. Thus, a set of 11 features is extracted with the purpose of reducing the dimension of data and hence reducing the computational effort. This set consists of the following features: standard deviation, entropy, Rényi entropy, Shannon entropy, mean deviation, Kurtosis, RMS value, crest factor, the balance between the maximum and minimum amplitude and peak value. Thus, for each dj data a Ck vector is extracted, where j represents the index of each element contained in the window and that varies in the range {1 → N} N is the size window; k represents each characteristic in the range {1 → 11}.

IV. RANDOM FOREST

Random Forest corresponds to a collection of combined Decision Tree {hk(x,Tk)}, where k = 1,2,...,L where L is number the tree and Tk is the training set built at random and identically distributed, hk represents the tree created from the vector Tk and is responsible for producing an output x.

Decision Trees are tools that use divide-and-conquer strategies as a form of learning by induction [5], Thus, this tool uses a tree representation, which helps in pattern classification in data sets, being hierarchically structured in a set of interconnected nodes. The internal nodes test an input attribute/feature in relation to a decision constant and, this way, determine what will be the next descending node. Therefore, the nodes considered as leaves classify the instances that reach them according to the associated label.

The trees that make up the Random Forest are built randomly selecting m (value fixed for all nodes) attributes in each node of the tree; where the best attribute is chosen to divide the node. The vector used for training each tree is obtained using random selection of the instances. Thus, to determine the class of an instance, all of the trees indicate an output, where the most voted is selected as the final result. So the classification error depends on the strength of individual trees of the forest and the correlation between any two trees in the forest.

V. RESULTS

As previously mentioned, the decision trees and the Random Forest were trained and validated using a set of data consisting of the signals windows acquired from the database. Thus, the training set is composed of 70% of the windows and the test/validation suite corresponds to the 30% remaining windows. The random forest is formed by 10 decision tree and the number of attributes selected in each node is equal to 4. This made it possible to obtain and evaluate the success rate for each disturbance, as well as the average accuracy of classifiers. The comparison of the classification results is presented in Table I.

TABLE I. RESULTS OBTAINED FOR SYNTHETIC SIGNALS .

Power Quality Disturbances	**DT**	**RFs**
Voltage Sags	83,0%	99,4%
Voltage Swells	94,4%	99,9%
Flickers	97,9%	99,9%
Harmonic Distortions	96,9%	99,6%
Voltage Sags with Harmonic Distortions	78,9%	98,5%
Voltage Swells with Harmonic Distortions	88,8%	98,9%
Voltage Interruptions	89,4%	99,3%
Oscillatory Transients	87,5%	99,2%
Mean Precision	**89,6%**	**99,3%**

Through the results presented in Table I it is found that the performance of the two used classifiers is satisfactory, however, it can be seen that the approach based on Random Forest presents better results when compared with the approach based on type J48 Decision Trees. The RF had a precision 10% higher than the DT, additionally the proposed algorithm can identify large part of the disturbances with an accuracy greater than 99%.

VI. CONCLUSIONS

The paper presents a performance comparison between type J48 Decision Trees and the Random Forest algorithm for classification of power quality disturbances. According to the results, we note that the worst performances were obtained for windowss containing combinations of voltage sags with harmonic distortion and swells with harmonic distortion (respectively, 98.5% and 98.9%). Therefore, in general, the results may be considered satisfactory for electric power systems.

REFERENCES

[1] R. C. Dugan, M. F. Mc Granagham, S. Santoso, and H. W. Beaty, Electrical Power Systems Quality, 3rd edition. New York, 2002.

[2] D. Granados-Lieberman, R. J. Romero-Troncoso, R. A. Osornio-Rios, A. Garcia-Perez, and E. Cabal-Yepez, "Techniques and Methodologies for Power Quality Analysis and Disturbances Classification in Power Systems: A Review," IET Generation, Transmission & Distribution, vol. 5, no. 4, pp. 519-529, 2011.

[3] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[4] H. Erişti, A. Uçar and Y. Demir, "Wavelet-based feature extraction and selection for classification of power system disturbances using support vector machines," Electric Power Systems Research, vol. 80, pp. 743-752, 2010.

[5] I. H. Witten and E. Frank, Data Mining: Pratical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005G.