We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

An Overview of Data Mining Techniques Applied to Power Systems

Jefferson Morais, Yomara Pires, Claudomir Cardoso and Aldebaro Klautau
Signal Processing Laboratory (LaPS) Federal University of Pará (UFPA)
Belém PA Brazil

1. Introduction

The growth of available data in the electric power industry motivates the adoption of data mining techniques. However, the companies in this area still face several difficulties to benefit from data mining. One of the reasons is that mining power systems data is an interdisciplinary task. Typically, electrical and computer engineers (or scientists) need to work together in order to achieve breakthroughs, interfacing power systems and data mining at a mature level of cooperation. Another reason is the lack of freely available and standardized benchmarks. Because of that, most previous research in this area used proprietary datasets, which makes difficult to compare algorithms and reproduce results.

This chapter has two mains goals and, consequently, is divided in two parts. In the first part, the goal is to present a brief overview on how data mining techniques have been used in power systems. There are several works, such as (Mori, 2002), that introduce data mining techniques to people with background in power systems. In contrast, this text assumes previous knowledge of data mining, describes some fundamental concepts of power systems and illustrates the kind of problems that the electric industry tries to solve with data mining.

The second part of the work presents a thorough investigation of a specific problem: classifying time series that represent short-circuit faults in transmission lines. Studies show that these faults are responsible for 70% of the disturbances and cascading blackouts (Kezunovic & Zhang, 2007). Besides, there is a large and growing number of publications about this problem.

Two types of fault classification systems are discussed: *on-line* and *post-fault*. On-line fault classification must be performed on a very short time span, with the analysis segment (or frame) being located approximately at the instant the fault begins. Post-fault classification can be performed off-line and its input consists of a multivariate time series with variable length (duration). Post-fault is a sequence classification problem, while in on-line classification the input is a fixed-length vector. Both fault classification systems (and most data mining applications) require a preprocessing or front end stage that converts the raw data into sensible parameters to feed the back end (in this case, the classifier).

Besides its practical importance, one reason for the popularity of fault classification is that it is relatively easy to artificially generate a dataset through simulators. Here, the well-known Alternative Transients Program (ATP) (ATP, 1987) simulator was used to create a public and

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

comprehensive labeled dataset. Such datasets are key components to allow reproducing research results in different sites, which is crucial given the large number of parameters to be tuned in a fault classification system. Therefore, in order to promote reproducible research, this work also provides detailed information about the adopted parameters.

The chapter is organized as follows. In Section 2 a brief description of data mining applications in power systems is provided. The section also introduces basic concepts of power systems. For a more detailed treatment, the reader is refereed to (Casazza & Delea, 2005). The second part begins in Section 3, which poses the fault classification problem and discusses solutions. Section 4 presents simulation results and is followed by the conclusions.

2. Power systems for data miners

2.1. Concepts of interest

A typical power system is represented in Figure 1 and can be divided into three parts: generation, transmission and distribution. The distribution system delivers power to the end users (*loads*). Most systems adopt three *phases* (A, B and C), using three conductors to carry sinusoidal voltage waveforms that have an offset in time equivalent to 120 degrees. While the customers need low voltage values (hundreds of Volts), the transmission system typically uses much higher values for efficiency. The transformers are responsible for the up and down conversions of voltage values and are located in different parts of the system.

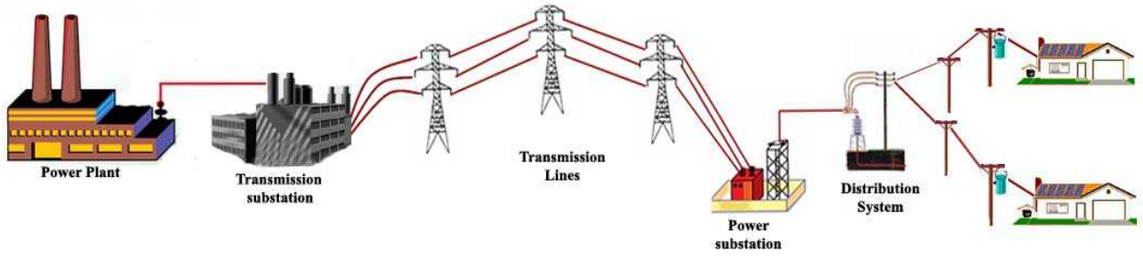


Fig. 1. An example of an electric power system.

Under normal conditions, the voltage waveform x(t) has a pre-established frequency (e.g., 60 or 50 Hz). Knowing the nominal value of the amplitude (e.g., 500 kV), it is convenient to normalize x(t) by this value and report the amplitude in p.u. (per unity). In this case, the ideal waveform could be expressed by $x(t) = \cos(2 \pi f t + \theta)$, where f is the frequency and θ the initial angle. Figure 2(a) illustrates a segment of the ideal voltage waveform for a frequency f = 60 Hz and θ = 0 radians. Figure 2(b) depicts simultaneously all three voltage waveforms in a segment containing a fault recorded by an *oscillograph* recording equipment: a short circuit between the conductor corresponding to phase B and ground (G). It can be seen that, besides phase B, the other two phases are also disturbed. Such faults belong to a category of events that is called *transients* because they tend to disappear after proper operation of the system to recover normal conditions, as occurs after approximately 0.05 seconds in Figure 2(b).

In order to properly operate it, a power system contains several data acquisition equipments. For example, some of these equipments register the status of logical (boolean) variables at each minute, while others store waveforms digitized at relatively high sampling

rates (e.g., 5 kHz). In many cases it suffices to monitor the root mean-square (RMS) value of each waveform estimated at each second. Figure 3 illustrates the information about the voltage amplitude that is provided by the RMS estimation.

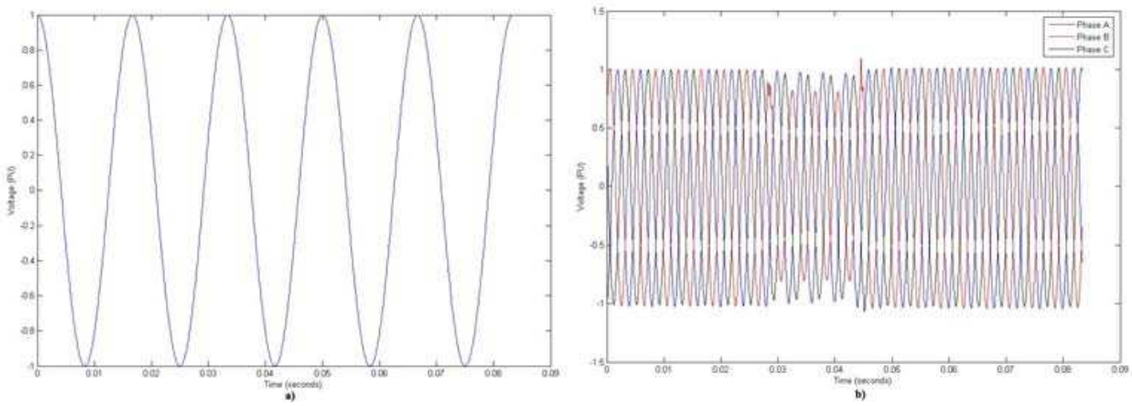


Fig. 2. a) Example of ideal normalized voltage waveform of one phase. b) All three phases with short-circuit between phase B and ground, as registered by an oscillograph equipment.

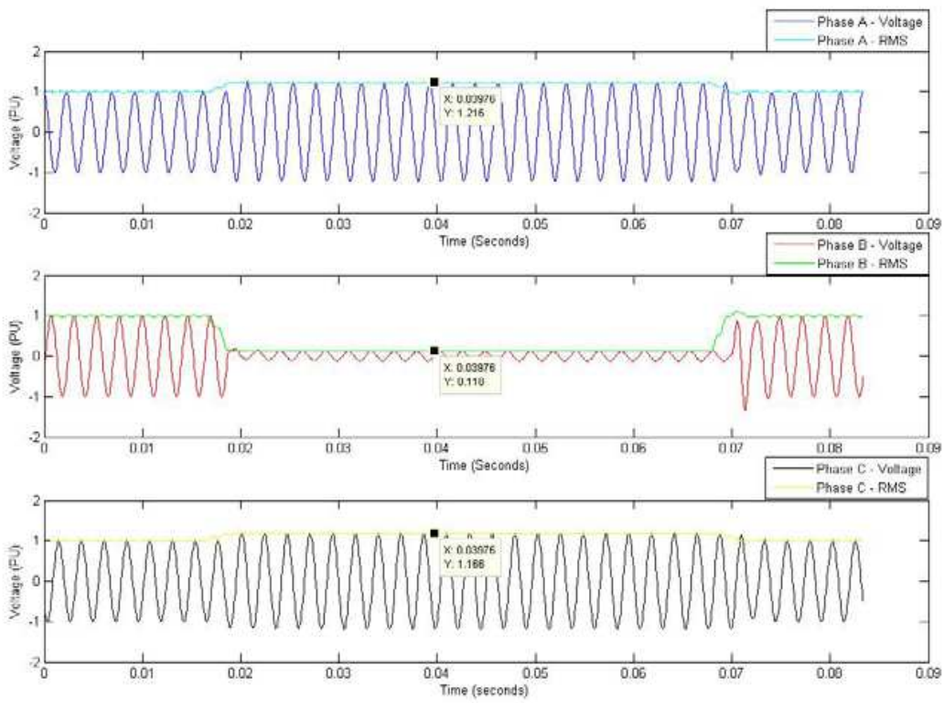


Fig. 3. Example of the RMS waveforms superimposed to the voltage waveforms in a fault between phase B and ground. It can be seen that the voltage in phase B drops to a value ower than 1 p.u., while phases A and C achieve values above the nominal.

In summary, most power systems data can be considered as multivariate time series, but the sampling frequencies may differ significantly and the variables are eventually categorical (not numerical). Roughly, one can organize the data originated in power systems into three categories:

a. Raw waveforms (voltages and currents) sampled at relatively high sampling frequencies;
b. Pre-processed waveforms (e.g., RMS) typically sampled at low sampling frequencies;
c. Status variables (e.g., if a relay is opened or closed) typically sampled at low sampling frequencies.

Due to its higher volume of information, the first category data is sometimes organized in specific databases, such as the oscillographic database, which stores all data from oscillographs. The other two categories are sometimes organized in the so-called Supervisory Control And Data Acquisition (SCADA) systems (Boyer, 1999). SCADAs are complex systems that periodically store several thousands of heterogeneous variables and are an important source of information for data mining. For example, automatically mining cause-effect relationships in SCADA data is an incipient but promising activity. Some power systems are affected by events that repeatedly cause troubles but their causes remain undetected. However, most of the times it is necessary to organize a data warehouse in order to be able to mine data from a SCADA, and fewer works use such data when compared to the first category.

Data mining can also alleviate another problem in power systems: when a disturbance is detected, a large amount of messages and alarms are generated. Protection equipments are responsible for detecting a problem, and act appropriately, isolating the defective part of system, for example. Part of this operation is automatic, but some tasks depend on a specialist. The amount of information regarding the problem cannot be excessive but should be enough for making decisions. Data mining techniques can be used to filter alarms and messages and provide the important information to the operator.

Failures in the performance of protection equipments, remote terminal, communications link and acquisition of date online, and variations in the voltage levels after of the occurrence disturbance, are factors that difficult the assessment and diagnosis in real time initial cause of power off.

Another problem of interest to the electric power industry is load forecasting, in which the goal is to predict the demand for power in specific regions. This can be cast as a conventional regression problem. Power quality is another area that can benefit from data mining. Here the goal is to help characterizing how close to the ideal (nominal) parameter values the system is operating. Small and large deviations are categorized by detection and classification modules. The location where a power quality event happened is also of interest.

The next section describes tasks and techniques used in data mining.

2.2 Review of data mining applications in power systems

This section briefly describes typical applications of data mining in electrical power systems via a collection of 18 papers. Table 1 lists the technique, task and application area.

Among the several applications listed in Table 1, the second part of this work concentrates in fault classification, as discussed in the next section.

Reference	Technique	Task	Application Area	Problem
Ramos, 2008	Decision tree	Classification	Distribution system	Characterization and classification of consumers
Hagh, 2007	Neural network	Classification	Transmission lines	Faults classification and locations
Saibal, 2008	WN[1]	Classification	Distribution system	Classification of transients
Chia-Hun & Chia-Hao 2006	Adaptative wavelet networks	Detection and discrimination	14-bus power system	Power-quality detection for power system disturbances
Pang & Ding 2008	wavelet transform and self-organizing learning array	Power quality disturbances classification	Distributed power system	Power-quality detection for power system disturbances
Bhende, 2008	Neural network	Classification	Not defined	Detection and classification of Power quality disturbances
Figueiredo, 2005	Decision tree	Classification	Distribution system	Electric energy consumer
Silva, 2006	Neural network	Detection and classification	Transmission lines	Faults detection and classification
Costa, 2006	Neural network	Classification	Transmission lines	Fault classification
Dola, 2005	Decision tree and neural network	Classification	Distribution system	Faults classification
Tso, 2004	Statistical analysis	Detection	Transmission and distribution systems	Detection the substations most sensitive to the disturbances
Mori, 2002	Regression tree and neural network	Forecasting	Distribution system	Load forecasting
Dash, 2007	Support Vector Machine	Classification identification	Transmission lines	Classification and identification of series-compensated

1Wavelet Networking (WN) can be considered as an extension of perceptron networks.

Monedero 2007	Neural network	Classification	Not defined	Classification of electrical disturbances in real time.
Vasilic, 2005	Fuzzy/ neural network	Classification	Transmission lines	Faults classification.
Vasilic, 2002	Neural network	Classification	Transmission lines	Faults classification.
Kezunovic, 2002	Neural network	Detection and diagnostic	Transmission lines	Detection and diagnosis of transient and faults.
Huisheng, 1998	Fuzzy/ neural network	Classification	Transmission lines	Faults classification.

Table 1. Summary of tasks, techniques and applications of data mining in power systems.

3. Classification of time series representing faults

As mentioned, most transmission systems use three **phases**: A, B and C. Hence, a short-circuit between phases A and B will be identified as "AB". Considering the possibility of a short-circuit to "ground" (G), the task is to classify a time series into one among ten possibilities: AG, BG, CG,AB, AC, BC, ABC, ABG, ACG and BCG. The ABC and ABCG faults are typically not distinguished because in well-balanced circuits (or ATP simulations) there is no current flow through the ground (Anderson, 1995). Algorithms to solve this classification problem are used by digital fault recorders (DFRs), distance relays and other equipments (Luo & Kezunovic, 2005).

The signal capturing equipments are typically located at both endpoints of transmission line. Most of them are capable of digitizing both voltage and current waveforms. It is assumed that a *trigger* circuit detects an anomaly and stores only the interval of interest: the fault and a pre-determined number of samples before and after the fault. The trigger is out of the scope of the present work and the simulations assumed a perfect trigger algorithm, with the fault endpoints being directly obtained from the simulator.

The next subsection describes the front end, the stage that is responsible for providing a suitable parametric representation of the time series. At some points the notation may look abusive, but there are many degrees of freedom when dealing with time series and a precise notation is necessary to avoid obscure points.

3.1 Front end

Each fault is a variable-duration multivariate time series. Then n-th fault $\mathbf{X}_n$ in dataset (oscillography records, for example) is represented by a $Q \times T_n$ matrix. A column of x_t, $t = 1,\ldots,T_n$, is a multidimensional sample represented by a vector of Q elements. For example, if we consider voltage and current waveforms of phases A, B and C, then $Q = \mathbf{6}$ in the experiments. In some situations, it is possible to obtain *synchronized samples* from both

endpoints of a given line. In these cases the sample is an augmented vector with twice the dimension of the single endpoint scenario. For the previous example, the sample dimension for double endpoint measures would be $Q = 12$.

A *front end* converts samples into *features* for further processing. An example of a modern front end algorithm is the wavelets decomposition (Vertelli & Kovacevic, 1995). Independent of the adopted parametric representation, a single sample typically does not carry enough information to allow performing reasonable decisions. Hence, it is useful to consider that a *front end* converts the matrix $\boldsymbol{X}$ in to a matrix $\boldsymbol{Z}$ with dimension $K \times N$, as depicted in Figure 4 (the processing is performed on $\boldsymbol{Z}$, not $\boldsymbol{X}$), where K is the number of features and N the number of feature vectors.

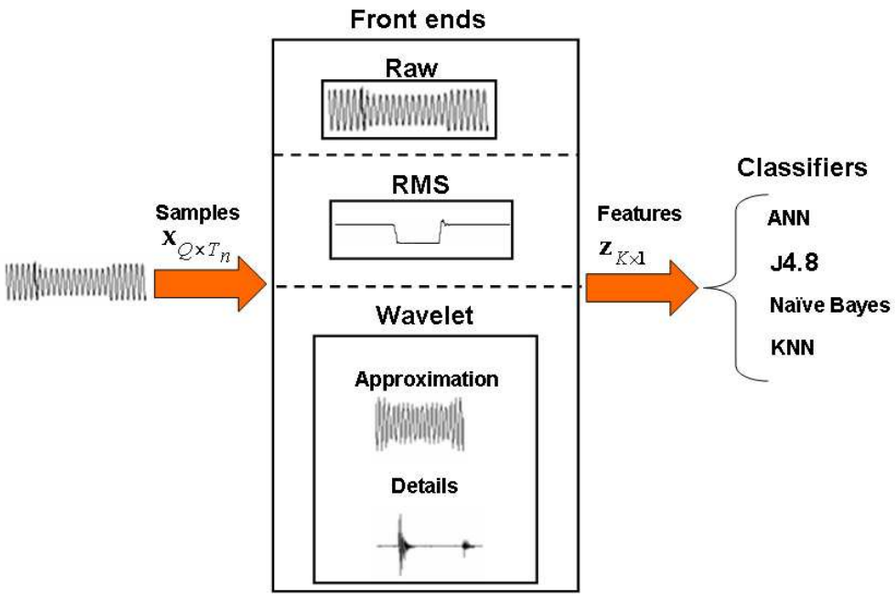


Fig. 4. The input and output matrices of the front end stage. Q and K are the dimensions of the sample and feature vectors, respectively, while T_n is the number of samples.

A *front end* is called raw when it outputs features that correspond to values of the original samples, without any processing other than organizing the samples into a matrix $\boldsymbol{Z}$. In the framed raw front end, this organization is obtained through an intermediate representation called *frame*. A frame $\mathbf{F}$ has dimension $Q \times L$, where L is the number of samples called *frame length* and their concatenation $\hat{\mathbf{z}} = [\mathbf{F}_1, \ldots, \mathbf{F}_n]$ is a matrix of dimension $Q \times LN$, where N is the number of frames.

The frames can overlap in time such that the *frame shift* S, i.e., the number of samples between two consecutive frames, is less than the frame length. Hence, the number of frames for a fault $\mathbf{X}_n$ is:

$$N_n = 1 + \left\lfloor (T_n - L)/S \right\rfloor \tag{1}$$

where $\lfloor \cdot \rfloor$ is the flooring function.

The frames $\mathbf{F}$ (matrices) are conveniently organized as vectors of dimension $K = QL$, and $\hat{\mathbf{Z}}$ resized to create $\mathbf{Z} = \lfloor \mathbf{z_1} \ldots \mathbf{z}_N \rfloor$ of dimension $K \times N$.

It should be noticed that, if $S = L$ (no overlap) and a frame is a *concatenation of samples*

$$\mathbf{F} = [\mathbf{x}_{t-0.5(L-1)}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+0.5(L-1)}] \tag{2}$$

the matrices $\mathbf{X}$ and $\hat{\mathbf{Z}}$ would coincide, i.e., $\hat{\mathbf{Z}} = \mathbf{X}$.

For example, in (Kezunovic & Zhang, 2007) the frames are composed by the concatenation of raw samples and vectors $\mathbf{Z}$ have dimension $K = \mathbf{198}$. In more details, for example, If $Q = \mathbf{6}$ (currents and voltages), a concatenated raw front end could obtain frames $\mathbf{F}$ of dimension 6×5 by concatenating to each central sample its four neighbours, two at the left and two at the right. In this case, assuming a fault with $T = 10$ samples and $S = L = 5$, one would have $K = 30$ and $N = 2$, such that $\hat{\mathbf{Z}} = \boldsymbol{X}$. In this case, $\hat{\boldsymbol{Z}}$ and $\boldsymbol{Z}$ would have dimensions 6×10 and 30×2, respectively. Figure 5 illustrates the segmentation in vectors $\mathbf{z}$ of features for one fault (ABG) with 4 frames. In this example, $L = 3$, $S = 1$ and this leads to three vectors $\mathbf{z}$, each of dimension $K = 18$.

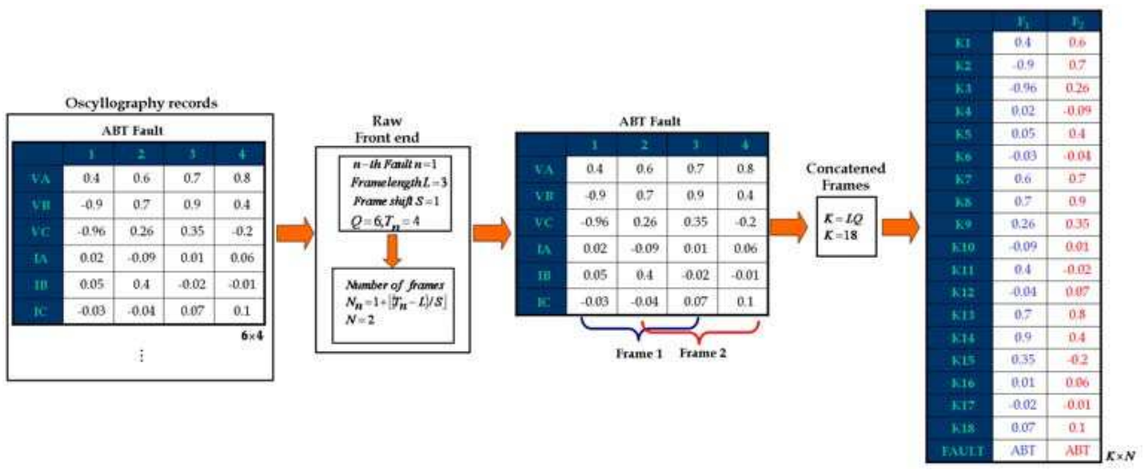


Fig. 5. Organizing feature vectors $\mathbf{z}$ in a concatenated raw front end. In this case, the ABG fault with a total of four frames, $L = 3$ and $S = 1$ lead to two vectors $\mathbf{z}$ of dimension $K = 18$.

As an alternative to the raw front end, the wavelet transform provides information via a multi-resolution analysis (MRA) (Vertelli & Kovacevic, 1995). When adopting this front end special care needs to be exercised to fully describe the processing, given their large number of degrees of freedom.

It is assumed a γ-level dyadic wavelet decomposition, which has γ stages of filtering and decimation (Vertelli & Kovacevic, 1995) and transforms each of the Q waveforms into γ+1 waveforms. More specifically, the q-th waveform is decomposed into approximation $\mathbf{a}^{\mathrm{q}}$ and details $\mathbf{d}_1^{\mathrm{q}}, \mathbf{d}_2^{\mathrm{q}}, \ldots, \mathbf{d}_\gamma^{\mathrm{q}}$, for $\mathrm{q} = 1, \ldots, Q$. For simplicity, the dependence on q is omitted hereafter.

Some works in the literature use only one of the details or calculate the average power of the coefficients (Morais et al., 2007). In contrast, the framed wavelet front end keeps all the

coefficients by taking in account that for $\gamma > 1$ they have different sampling frequencies and organizing them as matrix $\mathbf{Z}$. For that, instead of using a single L, the user specifies a value L_{min} for the waveforms with lowest f_s ($\mathbf{a}$ and $\mathbf{d}_\gamma$) and a large value $L_i = 2^{\gamma - i} L_{min}$, where S_{min} is another user-defined parameter.

The values are organized in a Frame $\mathbf{F}$ of dimension $Q \times L$, where $L = 2^\gamma L_{min}$. The number of frames for this organization of a wavelet decomposition is

$$N = 1 + \lfloor (T_a - L_{min}) / S_{min} \rfloor \quad (3)$$

where T_a is the number of elements in $\mathbf{a}$.

The notation is flexible enough to easily describe several wavelet front ends, such as the concatenated wavelet (*wavelet-concat*, for which $L_{min} = S_{min}$) and *wavelet-energy* described in (Morais et al, 2007).

Many recent works adopt the wavelet front end ((Saibal, 2008); (Silva, 2006); (Costa, 2006); (Chia-Hun & Chia-Hao, 2006); (Pang & Ding, 2008)). However, some of these works do not compare the wavelet with other (and eventually) simpler front ends. In Section 4, some results are presented with a simple RMS-based front end. It consists in taking the minimum RMS value of each phase during the whole fault duration. As a first step, a normalization is adopted (Morais et al., 2007) to to represent the voltage values in p.u. Because the feature vectors have dimension 3, it is relatively simple to visualize them. Figure 6 illustrates 1,000

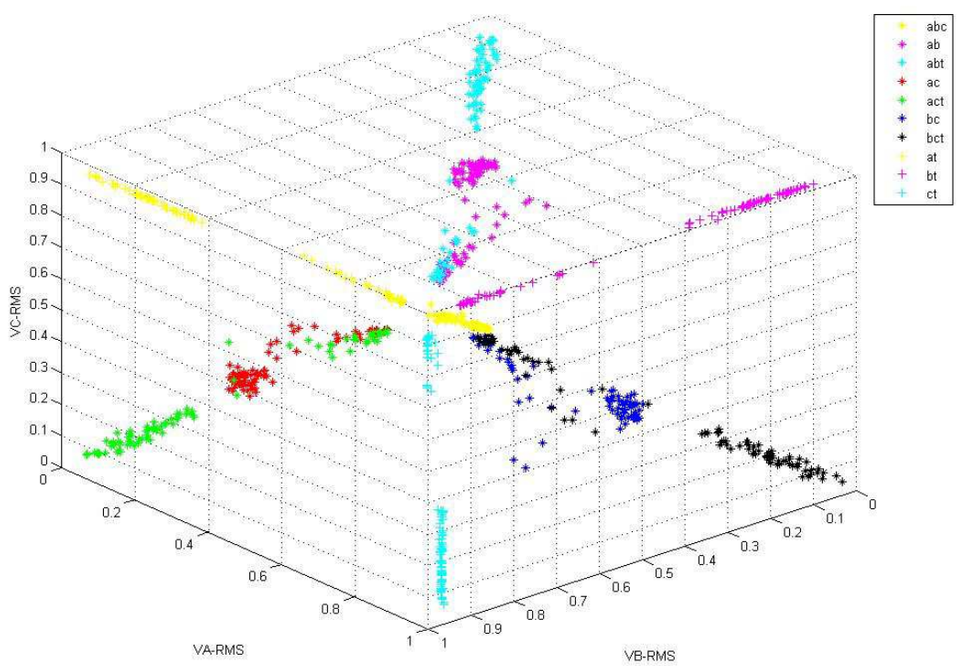


Fig. 6. Vector of features obtained with a simple RMS front end, which represents each fault by a three-elements vector. The color and shape indicate the fault category according to the legend.

of these vectors obtained via different simulations (more details in Section 4) with the color indicating the kind of fault. It can be seen, for example, that the monophasic faults (AG, BG e CG) and the triphasic fault (ABC) can be distinguished from the biphasics faults. Moreover, the biphasic faults that do not involve the ground are relatively similar to those biphasic faults involving ground.

3.2 On-line and post-fault classification

Fault classification systems can be divided into two types. The first one aims at performing a decision (classification) for each feature vector $\mathbf{z}$ or, equivalently, a frame $\mathbf{F}$ (giving that $\mathbf{z}$ is just a representation $\mathbf{F}$ as a vector). This is typically the goal in on-line scenarios, at the level of, e.g., a protection relay (Kezunovic & Zhang, 2007). Alternatively, the decision can be made at a supervisory center in a post-fault stage. The latter case makes a decision having available the whole matrix $\mathbf{Z}$ of variable dimension $K \times N_n$, where n distinguishes the individual faults, which having distinct durations in the general case. The on-line and post-fault systems try to solve problems that can be cast as *conventional classification* (Witten & Frank, 2005) and *sequence classification* (Ming & Sleep, 2005) problems, respectively.

On-line fault classification must be performed on a very short time span with the frame located in the beginning of the fault. It is often based on a frame corresponding to half or one cycle of the sinusoidal signal (typically of 60 or 50 Hz). For example, assuming 60 Hz and a sampling frequency of $f_S = 2$ kHz, one cycle corresponds to $L = 2000/60$ approximate 33 samples.

As mentioned, on-line classification corresponds to the conventional scenario, where one is given a *training* set $\{(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_M, y_M)\}$ containing M examples. Each example $(\mathbf{z}, y)$ consists of a vector $\mathbf{z} \in \Re^K$ called *instance* and a *label* $y \in \{1, \ldots, Y\}$. A conventional classifier is a mapping $\Phi : \Re^K \rightarrow \{1, \ldots, Y\}$. Some classifiers are able to provide *confidence-valued scores* $f_i(\mathbf{z})$ for each class $i = 1, \ldots, Y$ such as a probability distribution over y. For convenience, it is assumed that all classifiers return a vector $\mathbf{y}$ with Y elements. If the classifier does not naturally return confidence-valued scores, the vector $\mathbf{y}$ is created with a unitary score for the correct class $f_i(\mathbf{z}) = 1$ while the others are $f_i(\mathbf{z}) = 0$, $i \neq y$. With this assumption, the final decision is given by the *max-wins* rule.

$$F(\mathbf{z}) = \arg\max_i f_i(\mathbf{z}) \tag{4}$$

Contrasting to the on-line case, a post-fault module has to classify a sequence $\mathbf{Z}$. The classifier is then a mapping $\varphi : \Re^{K \times N} \rightarrow \{1, \ldots, Y\}$ and the training set $\{(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_M, y_M)\}$ contains M sequences and their labels. The technique adopted in this work is the *frame-based sequence classification* (FBSC) (Morais et al., 2007).

In FBSC systems, the fault module repeatedly invokes a conventional classifier $F(\mathbf{z})$ (e.g., a neural network or decesion tree) to obtain scores $\mathbf{y} = (f_1(\mathbf{z}), \ldots, f_Y(\mathbf{z}))$ for each class. To come up with the final decision, the fault module can then take in account the scores of all

frames. Two possible options consist in calculating an accumulated score $g_i(\mathrm{Z})$ for each class and then using the max-wins rule

$$G(\mathbf{Z}) = \arg\max_i g_i(\mathbf{Z}) \tag{5}$$

Where:

$$g_i(\mathbf{Z}) = \sum_{n=1}^{N} f_i(\mathbf{z}_n) \tag{6}$$

or

$$g_i(\mathbf{Z}) = \sum_{n=1}^{N} \log\left(f_i(\mathbf{z}_n)\right) \tag{7}$$

The accuracy of the system $G(\mathbf{Z})$ can be evaluated according to the misclassification rate and it is clearly dependent on the accuracy of the classifier $F(\mathbf{z})$. The misclassification rates are E_S and E_f, for the post-fault (sequence) and on-line (frame) modules, respectively. In the case of post-fault systems, in spite of E_S being the actual figure of merit, it is sometimes useful to also calculate E_f. However, one should note that estimating E_f takes in account all frames that compose a fault (frames in the beginning, middle and end of the fault). In on-line applications, such as relaying, taking a decision in the beginning of the fault is the most important. In order to take this situation in account, this work defines E_o as the misclassification rate obtained when one considers only the first frame of the fault.
The next section presents simulation results for fault classification.

4. Simulation results

The experiments used the UFPAFaults4 dataset, which can be downloaded from *www.laps.ufpa.br/freedatasets/UfpaFaults*. The UFPAFaults4 dataset is composed by 5,500 faults, organized into five sets of 100, 200, . . ., 1000 faults each. The division into these sets is to facilitate obtaining *sample complexity* curves (Vapnik, 1999). The sample complexity indicates how many training examples are required to train the classifier. It can be evaluated by observing how the performance varies with the number of training examples.
Each fault in the dataset corresponds to three voltage and three current waveforms stored as binary files with an associated text (ASCII) files, which stores a description of the fault (its endpoints, label, etc.). The waveform samples are stored as real numbers represented as the primitive type float in Java (big-endian, 32-bits, IEEE-754 numbers).
The faults are generated with the software AMAZONTP (Pires et al., 2005). Some parameters for the simulations are randomly generated. The values of all four resistances were obtained as i.i.d. samples draw from a uniform probability density function (pdf) U(0.1; 10), with support from 0.1 to 10 Ohms. The begin and duration (both in seconds) of the fault were draw from U(0.1; 0.9) and U(0.07; 0.5), respectively. The location was draw from U(2; 98) (percentage of the total line length). Eleven types of faults (AG, BG, CG, AB, AC, BC, ABC, ABG, ACG, BCG, ABCG) were generated using a uniform distribution.

The voltage and current waveforms generated by the ATP simulations had a sampling period equal to 0.25 microseconds, corresponding to a sampling frequency f_S = 40 kHz. It is possible to obtain versions with smaller values for f_S by decimating the original waveforms. This operation requires low-pass filtering to avoid aliasing. Details about decimation and filtering can be found in digital signal processing textbooks, e.g. (Oppenheim, 1989).

4.1 Normalization

The elements of feature vectors **z** may have very different dynamic ranges (e.g., voltage in kV and currents in Amperes). This can cause the learning algorithms to perform poorly. Therefore, as a pre-processing stage, it is important to apply a normalization process. There are many algorithms for normalization of time series. This work adopted the so-called *allfault* (Morais et al., 2007), which takes in account all duration of the waveforms for getting the maximum and minimum amplitudes of each phase, and the converting the values to the range [-1, 1]. A distinct normalization factor is calculated for each of the Q waveforms.

4.2 Model selection

Often, the best performance of a learning algorithm on a particular dataset can only be achieved by tedious parameter tuning. This task is called model selection and corresponds, for example, to choosing parameters such as the number of neurons in the hidden layer for a neural network. A popular strategy for model selection is cross-validation (Witten & Frank, 2005). This is a computationally intensive approach, but avoids tuning the parameters by repeatedly evaluating the classifier using the test set. The test set should be used only once, after model selection, such that the error rate on this test set is a good indicator of the generalization capability of the learning algorithm. When dealing with frames extracted from sequences, it should be noted that, in conventional classification, the examples are assumed to be i.i.d. "samples" from an unknown but fixed distribution $P(\mathbf{z}, y)$. Because examples are independent, they can be arbitrarily split into training and test sets. Similarly, when organizing the folds for cross-validation, examples can be arbitrarily assigned to the training and validation fold. However, the i.i.d. assumption becomes invalid, for example, when examples $(\mathbf{z}, y)$ are extracted from contiguous frames of the same sequence given the relatively high similarity among them. Hence, in practice it is important to use cross-validation properly, to avoid overtting due to a training set with similar vectors extracted from the same waveform.

This work performed model selection via a validation set, disjoint to both training and test sets. A grid (Cartesian product) of model parameters is created and the point (set of parameters) that leads to the smallest error in the validation set is selected. For each coordinate, the user specifies the minimum and maximum values, the number of values and chooses between a linear or logarithmic spacing for the values.

4.3 Results

The simulations in this work relied on Weka (Witten & Frank, 2005), which has many learning algorithms. Specifically, the work used decision trees (J4.8, which is a Java version of C4.5 (Witten & Frank, 2005)), multilayer artificial neural network (ANN) trained with backpropagation, naïve Bayes and K-nearest neighbor (KNN). The choice of these classifiers

was based in the fact that they are popular representatives of different learning paradigms (probabilistic, lazy, etc.).
The parameters obtained by model selection for each classifier are summarized in Table 2. The KNN used the squared-error as distance measure and K = 1 neighbors. The naïve Bayes used Gaussian pdfs and does not have parameters to be tuned. For the ANN, H is the number of neurons in the hidden layer, N the maximum the number of epochs, L the learning rate and M the momentum (Witten & Frank, 2005). For J4.8, C is the confidence and M the minimum number of examples in a leaf.

Front end	**L**	**S**	**K**	**ANN**	**J4.8**
Raw	1	1	6	-H 8 -N 1500 -L 0.2 -M 0.3	-C 0.35 -M 10
	5	5	30	-H 20 -N 1500 –L 0.2 -M 0.3	-C 0.5467 -M 10
	7	7	42	-H 26 -N 1500 –L 0.2 -M 0.3	-C 0.7433 -M 10
	9	9	54	-H 32 -N 1500 –L 0.2 -M 0.3	-C 0.35 -M 10
	11	11	66	-H 38 -N 1500 –L 0.2 -M 0.3	-C 0.5467 -M 10
	33	33	198	-H 104 -N 1500 -L 0.2 –M 0.3	-C 0.35 -M 10
RMS	33	33	3	-H 30 –N 2000 –L 0.1 –M 0.2	-C 0.35 –M 5

Table 2. Summary of parameters for the front ends and two classifiers.

The results for frame-based classification using the concatenated raw front end are shown in Figure 7**.** The best results were obtained by the ANN, followed by the J4.8 classifier. The best frame length was $L = 9$. It is interesting to note that for $L = 1$, ANN achieved an error rate more than three times the one achieved by J4.8, which could be due to problems in convergence.

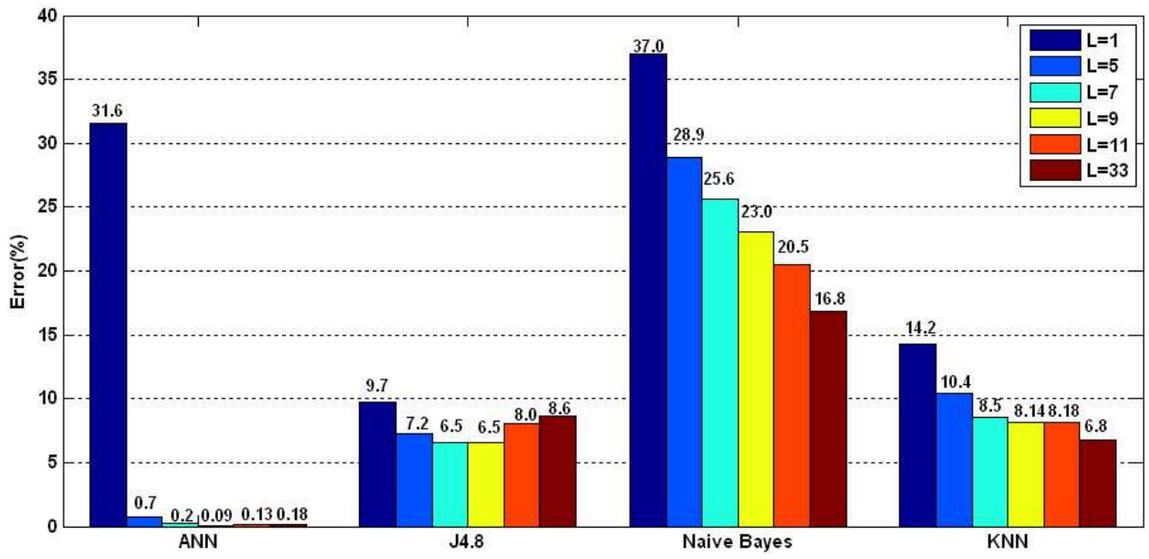


Fig. 7. Error rate E_f for several classifiers and frame lengths L using the concatenated raw front end.

The results obtained with RMS front end (Figure 8) were inferior to the ones in Figure 7, for the raw front ends. But it is interesting to note that the described RMS front end, which uses only three numbers obtained reasonable results.

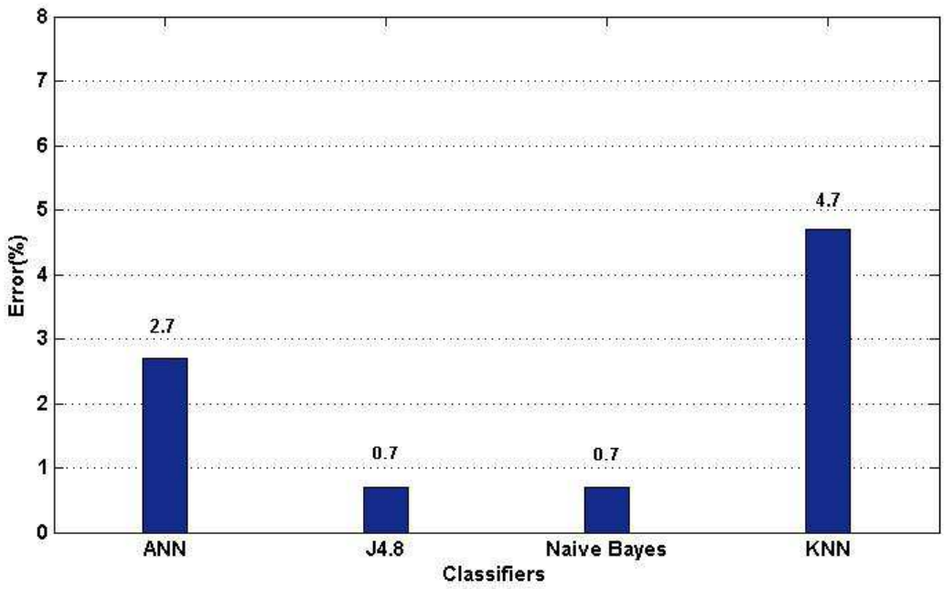


Fig. 8. Error rate E_S using RMS front end for several classifiers and frame length $L = 33.$.

Besides, the interpretation of decision trees trained with the RMS parameters provides a good insight about the problem. Figure 9 shows an example. In this case, the participation of a phase in the short circuit can be inferred by a relatively low minimum RMS value. For example, if this minimum value is above the threshold estimated from the data, then the corresponding phase should not be involved in the short-circuit.

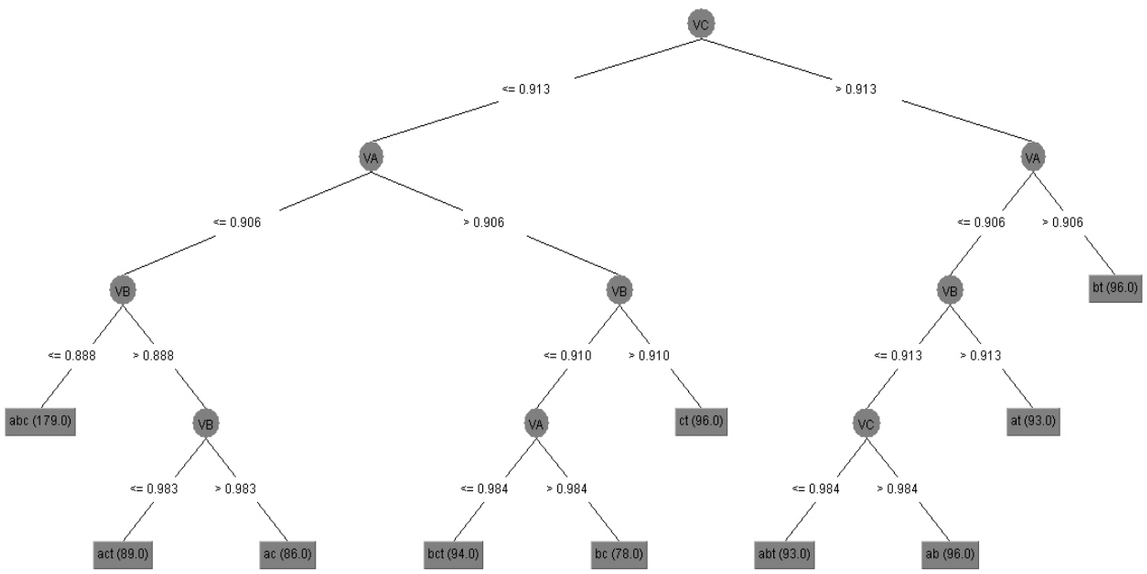
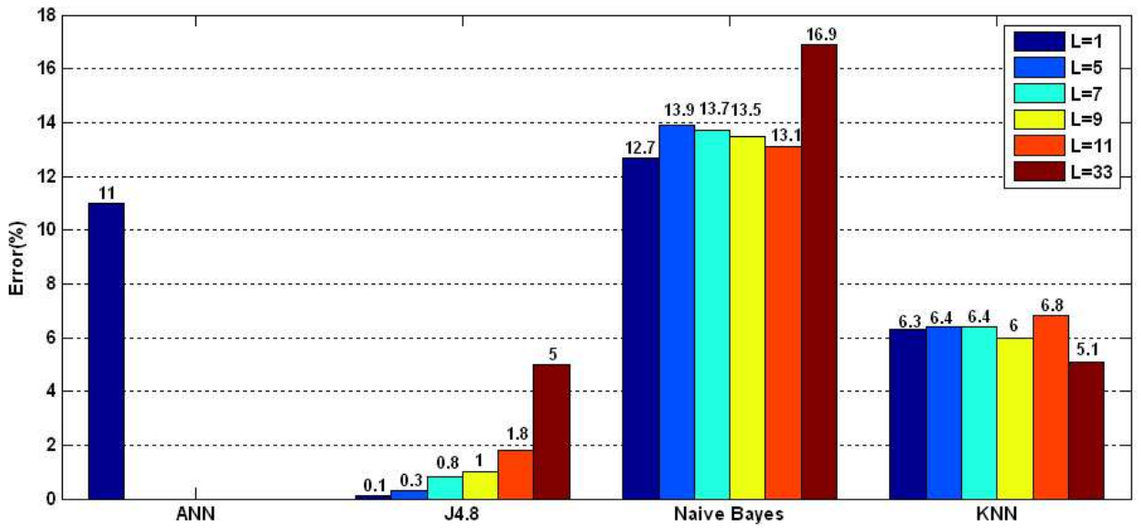
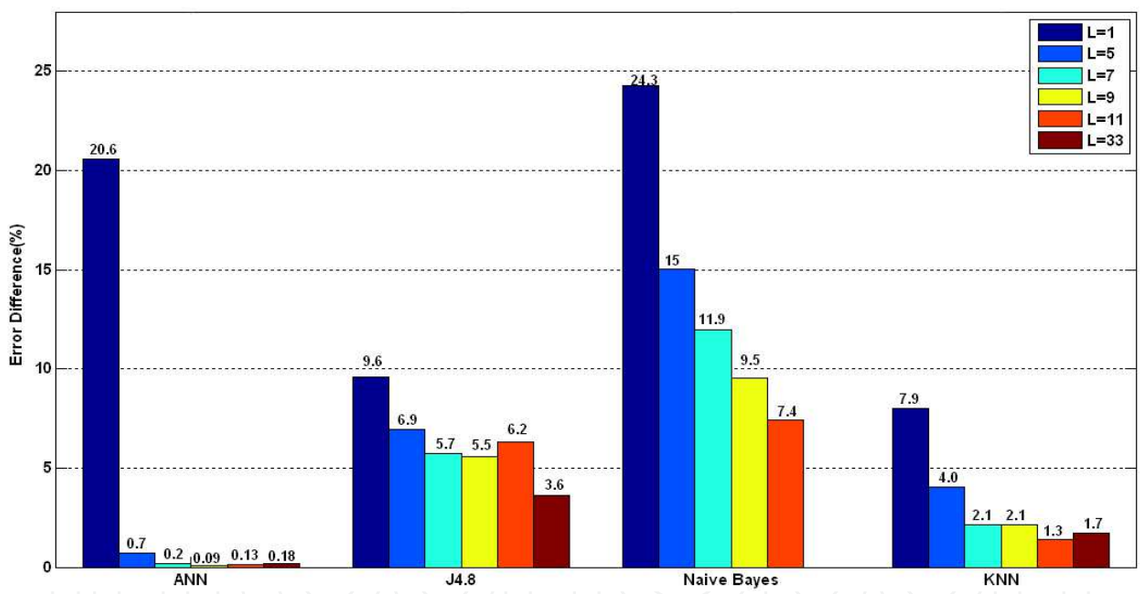


Fig. 9. Decision tree for the RMS front end that represents each fault with only three parameters: VA, VB and VC, which correspond to the minimum RMS value of each phase.



a)



b)

Fig. 10. Results for post-fault classification. a) Error E_S for post-fault classification. The ANN-based FBSC achieved $E_S = 0$ for $L > 1$. b) Difference $E_f - E_S$ between the erro rates for frame-by-frame and sequence classification.

Figure 10 shows results for post-fault classification. Figure 10 a) shows absolute values while Figure 10b) indicates the difference between E_S and E_f. As expected, post-fault classifiers can achieve smaller error rates than the ones that operate at one frame only. One can see that the ANN-based FBSC achieves $E_s = 0\%$ for all values of L but $L = 1$. The classifier J4.8 achieves $E_s = 0.1\%$ with a computational cost smaller than the ANN.

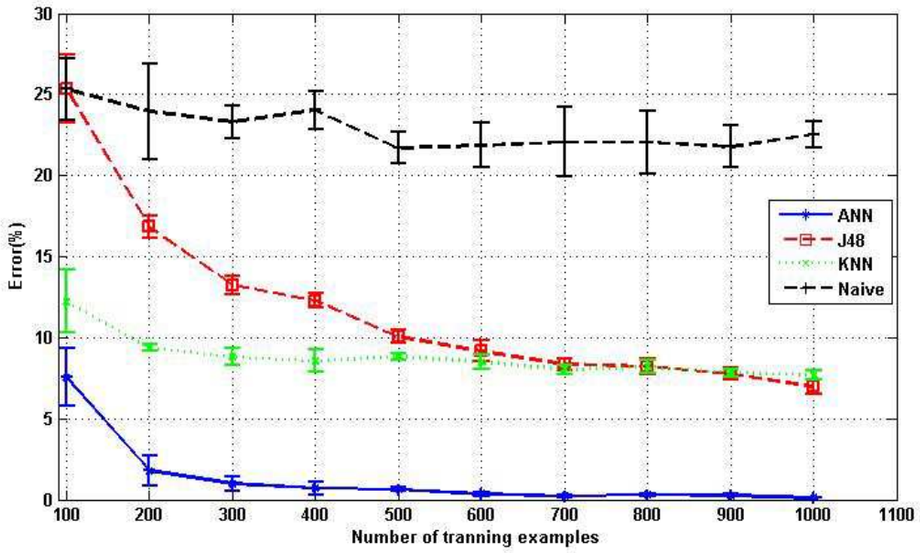


Fig. 11. Sample complexity for frame-based classification (the error is E_f) with $L = S = 9$. The figure shows the average and standard deviation. It can be seen that approximately $M = 700$ examples suffices to train the classifiers.

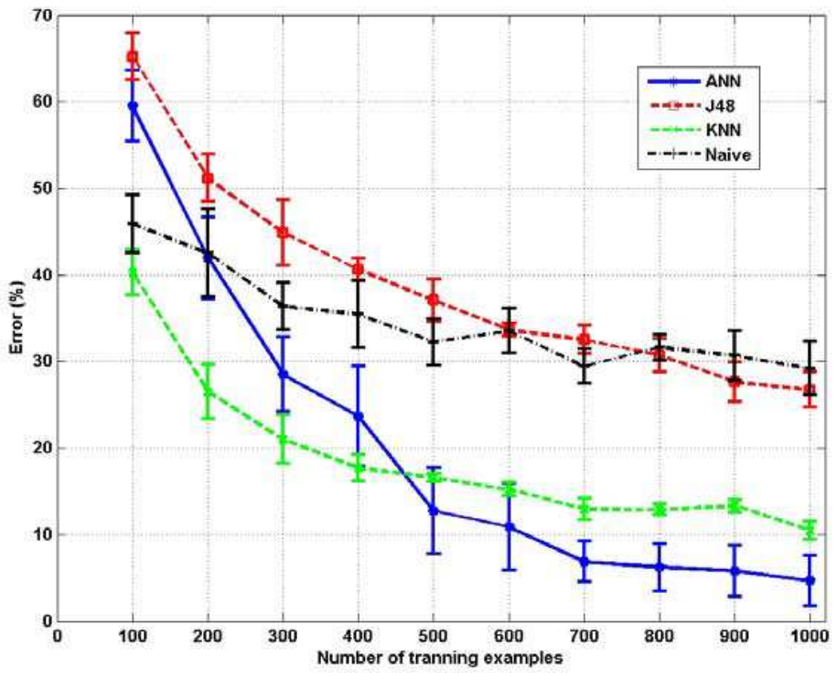


Fig. 12. Sample-complexity for one-frame-based classification with $L = S = 9$ (error E_o). The figure shows the mean and standard deviation. It can be seen that the error has a more erratic behavior than in Figure 11.

Figures 11 and 12 show, respectively, the results for sample complexity of frame-by-frame E_f and one-frame E_0 classification with $L = S = 9$. Model selection was used for each value of M, given that the best parameters for the classifier typically depend on the number of training examples (Rifkin & Klautau, 2004) . Comparing Figures 11 and 12 one can conclude that more examples are needed to train a classifier that observes only the first frame of the fault and its misclassification rate E_0 is typically higher than E_f under the simulated conditions.

5. Conclusions

This work presented an overview of data mining techniques used in power systems. Among several data mining tasks, fault classification is popular especially because it is relatively easy to generate artificial data using simulators such as ATP. Other applications of data mining will potentially impact the electric power industry, but this will require data warehouses to cope with preprocessing and organizing heterogeneous and large datasets.

Even within fault classification, the research methodology needs improvement for an easier conversion of academic results into effective products. One important issue is the robustness of proposed algorithms to distinct power system constituent elements, such as the transmission line lengths. Several algorithms are tested with only one simulated scenario. This work shows that a very simple RMS front end, which represents each fault by only three values, can lead to misclassification rates under 1% in controlled conditions. Hence, it is important to improve benchmarks with publicly available datasets, such as the UFPAFaults4, and use them to properly evaluate new approaches.

This chapter also presented results comparing different figures of merit for evaluating fault classification systems. It was shown that post-fault classifiers, which can take the whole fault segment in account to make a decision, achieve smaller error rates than classifiers based on a fixed-length (and short) frame. In fact, neural networks precisely classified all test examples (zero error) in some configurations. Another aspect that was emphasized is that, as vastly discussed in the machine learning literature, the number of examples to train a classifier depends on the learning algorithm and the domain (dataset).

6. References

Mori, H.; Kosemura, N.; Kondo, T.; Numa, K.; (2002). Data mining for short-term load forecasting. *Power Engineering Society Winter Meeting*. pp. 623 – 624, ISBN 0-7803-7322-7, Jan 2002.

Kezunovic, M.; Zhang, N. (2007). A real time fault analysis tool for monitoring operation of transmission line protective relay, *Electric Power Systems Research*, Vol.77, No.3-4, (March 2007) 361-70.

ATP, (1987). Alternative Transients Program, Rule Book, Leuven EMTP Center (LEC).

Casazza, J.; Delea, F. (2005). *Understanding Electric Power Systems - An Overview of the Technology and the Marketplace*. Electrical Insulation Magazine, IEEE. ISBN: 978-0-471-44652-1.

Boyer, Stuart A.(1999). SCADA: Supervisory Control and Data Acquisition , 2nd Edition, ISA International Society for Measurement. ISBN 1-55617-660-0.

Ramos, S.; Vale, Z. (2008). Data mining techniques application in power distribution utilities. *Transmission and Distribution Conference and Exposition,* pp. 1-8. ISBN. 978-1-4244-1903-6, Chicago, April 2008.

Hagh, M.T.; Razi, K.; Taghizadeh, H. (2007). Fault classification and location of power transmission lines using artificial neural network, *International Power Engineering Conference*, pp. 1109 – 1114, ISBN 978-981-05-9423-7, Singapore, Dec 2007.

Saibal Chatterjee, Sivaji Chakravorti, Chinmoy Kanti Roy and Debangshu Dey. (2008) Wavelet network-based classification of transients using dominant frequency signature, *Electric Power Systems Research,* Vol. 78, No. 1, (January 2008) 21-29.

Chia-Hung Lin; Chia-Hao Wang.(2006).Adaptive wavelet networks for power-quality detection and discrimination in a power system, *IEEE Transactions on Power Delivery*, Vol 21, No. 3,(July 2006) 1106 - 1113, ISSN: 0885-897.

Pang, P.; Ding, G.(2008). Power quality detection and discrimination in distributed power system based on wavelet transform. *27th Chinese Control Conference (CCC 2008)*, pp. 635 – 638, ISBN 978-7-900719-70-6, China, July 2008.

Bhende C. N.; Mishras S.; Panigrahi B. K. (2008). Detection and classification of power quality disturbances using S-transform and modular neural network . *Electric Power Systems Research*, Vol. 78, (February 2008) 122-128.

Figueredo, V.; Rodrigues F.; Vale, Z.; Gouveia, J. B. (2005). An eletric energy Consumer characterization Framework based on data mining techniques. *IEEE Transactions Power Systems*, Vol. 20, No. 2., May 2005 , 596- 60), ISSN 1558-0679.

Silva, K. M.; Souza, B. A.; Brito, N. S. D. (2006). Fault detection and classification in transmission lines based wavelet transform and ANN. *IEEE Transaction on Power Delivery*, Vol 21 , No. 4, (October 2006) 2058-2063, ISSN 0885-8977.

Costa, F. B.; Silva, K. M.; Souza, B. A.; Dantas, K. M. C.; Brito, N. S. D. (2006). A method for fault classification in transmission lines bases on ANN and wavelet coefficients energy. *International Joint Conference Neural Networks*. pp. 3700 – 3705, ISBN 0-7803-9490-9, Vancouver, July-2006.

Dola, H.M.; Chowdhury, B.H. (2005). Data mining for distribution system fault classification. *Power Symposium, 2005. Proceedings of the 37th Annual North American*, pp. 457 – 462, ISBN 0-7803-9255-8, October 2005.

Tso, S.K.; Lin, J.K.; Ho, H.K.; Mak, C.M.; Yung, K.M.; Ho, Y.K. (2004). Data mining for detection of sensitive buses and influential buses in a power system subjected to disturbances. *IEEE Transactions on Power Systems*, Vol. 19, No.1, (February 2004) 563 – 568, ISSN 1558-0679.

Dash, P.K.; Samantaray, S.R.; Panda, G. (2007). Fault Classification and Section Identification of an Advanced Series-Compensated Transmission Line Using Support Vector Machine. *IEEE Transactions on Power Delivery*, Vol 22. No. 22, (January 2007) 67 – 73, ISSN 0885-8977.

Monedero, I.; Leon, C.; Ropero, J.; Garcia, A.; Elena, J.M.; Montano, J. C. (2007). Classification of Electrical Disturbances in Real Time Using Neural Networks. *IEEE Transactions on Power Delivery*, Vol. 22, No. 3, (July 2007) 1288 – 1296, ISSN 0885-8977.

Vasilic, S.; Kezunovic, M. (2005) Fuzzy ART Neural Network Algorithm for Classifying the Power System Faults. *IEEE Transactions on Power Delivery,* Vol. 20, No. 2, (April 2005) 1306-1314, ISSN 0885-8977.

Vasilic, S.; Kezunovic, M. (2002). An Improved Neural Network Algorithm for Classifying the Transmission Line Faults. *IEEE Power Engineering Society Winter Meeting*, pp. 918 – 923. ISBN 0-7803-7322-7, Jan 2002.

Kezunovic, M. ; Vasilic, S.; Gul-Bagriyanik, F. (2002). Advanced Approaches for Detecting and Diagnosing Transients and Faults. 2002.

Huisheng Wang; Keerthipala, W.W.L. (1998). Fuzzy-neuro approach to fault classification for transmission line protection. *IEEE Transactions Power Delivery*, Vol.13, No. 4, 1093-1104, ISSN 0885-8977.

Anderson, P. M. (1995). *Analysis of faulted Power Systems*. IEEE Press Series on Power Engineering, ISBN 978-0-7803-11459.

Luo, X.; Kezunovic, M.(2005). Fault Analysis Based on Integration of Digital Relay and DFR. *Power Engineering Society General Meeting,* pp. 746 – 751, ISBN 0-7803-9157-8, Jun 2005.

Vertelli, M.; Kovacevic, J. (1995). *Wavelets and Subband Coding*. Prentice Hall, ISBN 978-0130970800.

Morais J.; Pires, Y.; Cardoso, C.; Klautau, A. (2007). Data mining applied to the electric power industry: Classification of short-circuit faults in transmission lines. *In IEEE International Conference on Intelligent Systems Design and Applications* (ISDA), pp. 943-948, ISBN: 978-0-7695-2976-9 , October 2007.

Witten, I.; Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2nd Edition, ISBN 0-12-088407-0.

Ming Li.; Sleep, R. (2005). A robust approach to sequence classification, *International Conference on Tools with Artificial Intelligence*, pp.5, ISBN 0-7695-2488-5, November 2005.

Vapnik, V.N (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, Vol 10, No.5, Sept. 1999.988 – 999, ISSN: 1045-9227.

Pires, Y. P.; Santos, A.; Borges, J.; Carvalho, A.; Nunes, M. V. A.; Santoso, S.; Klautau, A. (2005). A framework for evaluating data mining techniques applied to power quality. *Brasilizn Conference on Neural Network*, October 2005.

Yang, K.; Shahabi, C. (2004). A PCA-based similarity measure for multivariate time series. 2nd, ACM International Workshop on Multimedia DataBases, pp. 65-74, ISBN 1-58113-975-6, Washington, DC, USA.

Oppenheim, A.; Schafer, R. (1989). *Discrete-time Signal Processing*. Prentice-Hall, 2nd Edition , ISBN 9780137549207.

Rifkin, R.; Klautau, A. (2004). *In defense of one-vs-all classification*. Journal of Machine Learning Research, 5:101-141, 2004.

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jefferson Morais, Yomara Pires, Claudomir Cardoso and Aldebaro Klautau (2009). An Overview of Data Mining Techniques Applied to Power Systems, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/an_overview_of_data_mining_techniques_applied_to_power_systems

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen