Rev. Fac. Ing. Univ. Antioquia N.° 44. pp. 83-96. Junio, 2008

Classification methodology and feature selection to assist fault location in power distribution systems

Metodología de clasificación y de selección de atributos para localización de fallas en sistemas de distribución de energía eléctrica

Juan José Mora Flórez, Germán Morales España, Sandra Pérez Londoño*

Grupo de Investigación en Calidad de Energía Eléctrica y Estabilidad (ICE) - Programa de Ingeniería Eléctrica, 227. Universidad Tecnológica de Pereira. Pereira, Risaralda, Colombia.

(Recibido el 9 de octubre de 2007. Aceptado el 29 de enero de 2008)

Abstract

A classification methodology based on Support Vector Machines (SVM) is proposed to locate the faulted zone in power distribution networks. The goal is to reduce the multiple-estimation problem inherent in those methods that use single end measures (in the substation) to estimate the fault location in radial systems. A selection of features or descriptors obtained from voltages and currents measured in the substation are analyzed and used as input of the SVM classifier. Performance of the fault locator having several combinations of these features has been evaluated according to its capability to discriminate between faults in different zones but located at similar distance. An application example illustrates the precision, to locate the faulted zone, obtained with the proposed methodology in simulated framework. The proposal provides appropriate information for the prevention and opportune attention of faults, requires minimum investment and overcomes the multiple-estimation problem of the classic impedance based methods.

---------- *Keywords:* distribution systems, fault location, power quality, signal characterization, support vectors.

* Autor de correspondencia: teléfono: +57 +6 +321 17 57, fax: +57 +6 +313 71 22, correo electrónico: jjmora@utp.edu.co (J. Mora-Flórez).

Resumen

En este artículo se propone una metodología de clasificación basada en máquinas de soporte vectorial (SVM), para localización de fallas en redes de distribución de energía eléctrica. La meta es reducir el problema de la estimación múltiple de los métodos que usan medidas en un sólo terminal de la línea (la subestación), para estimar la localización de fallas en sistemas radiales. Adicionalmente, la selección de características o descriptores obtenidos de la señal de tensión y de corriente se analiza y usa como entrada del clasificador SVM. Se evalúa además el desempeño del localizador ante diferentes combinaciones de estas características, de acuerdo con su capacidad para discriminar entre las fallas que ocurren en la diferentes zonas del sistema de potencia, pero localizadas a una distancia similar desde el punto de medida. Un ejemplo de aplicación ilustra la precisión para localizar la zona en falla obtenida con la metodología propuesta y en un entorno de simulación. El localizador propuesto suministra información para la prevención y atención oportuna de fallas, requiere de mínima inversión y soluciona la múltiple estimación de los métodos basados en la estimación de la impedancia.

---------- *Palabras clave*: Sistemas de distribución, localización de fallas, calidad de potencia, caracterización de señales, vectores de soporte.

Introduction

Faults in power distribution systems cause supply interruptions being responsible of process disturbances, information and economic loss and equipment damage among others [1, 2]. High accuracy approaches have been proposed for fault location in power transmission systems; however, these algorithms are not useful for fault location in distribution systems due to the specific characteristics of the latter [3, 4, 5]: a) Single end measurements of voltage and current are available only at the distribution substation; b) Distribution systems are usually operated in a radial mode and they are characterized by the presence of single and double phase laterals; c) Loads are usually tapped along the lines and laterals and could be either single or multi-phase; and, d) The non-uniform development of the network and variations on loads are responsible of heterogeneous section lines (presence of different conductor gauges, combination of overhead lines and underground cables, etc.).

Different methods have been proposed for fault location in power distribution systems. Most of them are based on the calculation of the equivalent impedance as seen from the substation during the fault. The pre-fault and fault effective values (rms) of the fundamental current and voltage at the substation are used with this purpose [5, 6]. Then, the faulted section is estimated following an iterative procedure: the impedance obtained from the line model considering a possible fault in each node of the network is compared with the equivalent impedance calculated from measurements. The fault point in the section line could be estimated according to the reactance analysis as it is described in [6]. There exist a variety of methods that follows this basic principle for fault location in distribution systems. Thus, the effect of loads and laterals is studied in references [7] and [8] and the complexity of dealing with heterogeneous lines is analyzed in reference [9]. Algorithmic solutions have been proposed in the literature to deal with those considerations with good accuracy results, as presented by Das [5], Saha [10] and Son Choi et al. [11]. The main drawback of these methods is the multiple-estimation problem given by the existence of multiple points (usually far away one from each other) in the power distribution system that fulfill with the equivalent impedance condition. Consequently, these methods provide precise (accurate distance to the fault) but uncertain (multiple sections are at the same electrical distance) fault locations. This is the main disadvantage to apply these methods in real systems since the existence of multiple possible fault locations in a large geographical area do not solve the need of a quick fault location and introduce a decision problem: which is the most likely faulted node?

Recently, many works have addressed this problem by using knowledge-based techniques to exploit the existence of previous experiences and contextual information. In reference [12] current measured at the power substation have been used to train an ANFIS net to associate current patterns with protective device settings in order to isolate faulted zones in the system. In a similar strategy, heuristic knowledge from operators is exploited together with information from the SCADA database [13]. Fuzzy inference is used in both methods to deal with uncertainty inherent in these methods. A Bayesian network has been proposed [14] as a causal model between the fault equipment and the evidences of observations during feeder outages as the regional distribution of trouble calls or abnormal observations of the feeders and surrounding environments expressed in the calls. A rule-based expert system is described [15] to locate the faults in a distribution system by using component data and network topology stored in the database and a set of rules defined by engineers. Data mining methodology have been used [16] to derive fault location and diagnosis models from reports containing similar information. A database containing information from customers and SCADA system during outages is used [17] to reduce the search area associated to faults. In this method a second step is proposed for fault location assisted by the design of a meter-polling schema based on the existence of a complete automated meter reading system. All of the previous referenced methodologies use information from a database of equipment faults, SCADA, customer calls among others, not always available in most of the distribution facilities.

In this paper, we propose an alternative method that only uses information contained in the single end measurements of current and voltage at the distribution substation. A procedure to reduce the multiple-estimation problem is proposed based on the use of significant features extracted from voltage and current registers and the use of Support Vector Machines (SVM) as classification tool. Non-linear classification properties of SVM have been exploited to assign fault registers with predefined fault sections. Features, extracted from voltage and currents, have been selected to be sensible to fault location and fault type. The following features have been considered with this purpose: Variation of the rms values of voltage, current and apparent power, the reactance seen from the substation during the steady state of fault and finally the frequency of the transient caused by the fault.

This paper is presented in six sections. Section two is devoted to introduce the fundamentals of SVM. In section three, the definition and the procedure used to characterize the voltage and current signals to be used in fault location is presented. The fault location strategy is described in section four. Section five exemplifies this proposal in a real power distribution system. Finally, section six is devoted to conclude and summaries the contributions of this work.

Support Vector Machines (SVM)

Support Vector Machines (SVM) have been used in this work as classification technique to assist fault location because the good results reported in diagnosis applications. SVM are based on statistical learning theory and can be viewed for practical purposes as a binary classification technique resulting from the development of artificial neural networks and its combination with the optimization and generalization theories [18, 19]. In the following subsections a brief summary of the theory is presented in three steps: the case of dealing with linear separable data, the use of soft margin constrain to deal with noisy data and finally the use of kernel functions in the non linear separable case.

Linear case

Suppose having *n* training elements, $\boldsymbol{x}_i$, in an $\boldsymbol{N}$ dimensional space. Each element has its respective *tag etiquette(y)* as presented in (1). This etiquette is used to label members of the same class (+1 or -1).

$$\boldsymbol{x}_i \in R^N \text{ and } \boldsymbol{y}_i \in \{+1,-1\} \tag{1}$$

Training data is used to estimate a decision function, which causes an output in {±1} when it is excited with an input $\boldsymbol{x}_i$ in R^N. The goal is to find a hyper plane *H*: $y=\boldsymbol{w}\cdot\boldsymbol{x}-b=0$ and two parallel hyper planes, *H1:* $y=\boldsymbol{w}\cdot\boldsymbol{x}-b=+1$ and *H2:* $y=\boldsymbol{w}\cdot\boldsymbol{x}-b=-1$, conditioned by the inexistence of elements between *H1* and *H2* and forcing the distance between them (margin) to be maximum ($\boldsymbol{wx}$ represents the inner product between vector $\boldsymbol{w}$ and $\boldsymbol{x}$). The hyper plane *H* is designed as *optimum separating hyper plane* (OSH). Figure 1 illustrates this requirement in a two-dimension space.

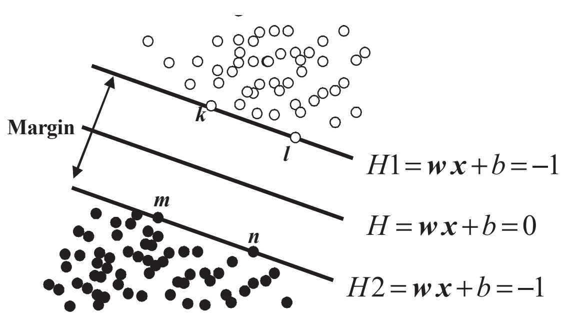

Figure 1 Separating hyper planes

Weight ($\boldsymbol{w}$) and bias (b) are the only parameters used to control the function and those data points that the margin pushes up against are called *Support Vectors*.

To find the OSH it is necessary to find those Support Vectors or in other words to perform margin maximization by solving the objective function presented in (2) (Margin is inversely proportional to $\sqrt{\boldsymbol{w}\cdot\boldsymbol{w}}$)

Subject to $$\min_{w,b} \frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w}) \quad (2)$$

$$\boldsymbol{y}_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 \ \forall i$$

This is a typical linear constrained quadratic optimization problem, convex in a convex set (***w**, b*). Therefore, it has only one possible solution. When Lagrange multipliers, with $\alpha_i \geq 0$, are used to solve this type of problems the problem is transformed to equation (3).

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b)\right] \quad (3)$$

Using this dual representation (3) the solution is obtained by performing a minimization of *L*. A later maximization in α_i allows finding the saddle point. The points with α_i different from zero correspond to the Support Vectors (SVs). These are the critical elements in the original dataset in the sense that when the training process is performed taking only into account these SVs, the same hyper planes will be obtained. Figure 1, shows the SVs data points, labeled as *k*, *l*, *m* and *n*.

Soft margin

The methodology presented in the previous sections is based on the no existence of mixed classes. In order to cope with this common problem the previous strategy is reformulated by considering a relaxation in the optimization condition to define what is known as a "*soft margin*". In order to allow not satisfying completely the function constrains presented in (2), the slack variables (ξ_i) are introduced and a new set of constrains is formulated as in (4).

Subject to $$\boldsymbol{y}_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i \ , \forall_i \quad (4)$$

$$\xi_i \geq 0, \forall i$$

Figure 2 shows this situation with nonlinearly separable classes. The slack variables ξ_i allow measuring the classification error in terms of distance to the hyper plane.

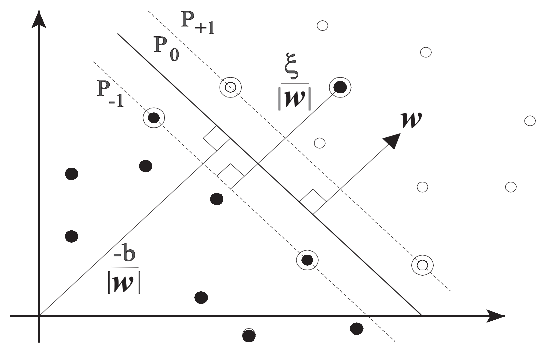

Figure 2 Non-separable case by using a linear hyperplane

A classifier can be obtained by controlling its classification capability using ||w|| and the criteria related to the slack variables that penalize the training errors. Typically the number of training errors (5) is used with this aim.

$$\left(\sum_{i=1}^{n} \xi_i\right) \quad (5)$$

And the optimum hyper plane can be obtained by the soft margin criteria in (6).

Subject to $$\min_{w,b} \frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w}) + C\sum_{i=1}^{n}\xi_i \quad (6)$$

$$\boldsymbol{y}_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i \ , \forall_i$$

Parameter *C* is denoted as "error penalization constant", and has to be fixed a priory by the user. High value of *C* means a high penalization error. When the Lagrange multipliers are used, the problem is transformed in (7).

Subject to $$\max_{\alpha} \sum_{i=1}^{n}\boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i,j=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\boldsymbol{y}_i\boldsymbol{y}_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \quad (7)$$

$$0 \leq \boldsymbol{\alpha}_i \leq C, \forall i \ \textit{and} \ \sum_{i=1}^{n}\boldsymbol{\alpha}_i\, y_i = 0$$

Kernel based SVMs

In case of non-linear separable dataset, it is possible to transform the data set to a new dimensional space, where the data is linearly separable. Figure 3 presents the intuitive idea of this transformation. The transformation function, $\Phi(.)$, is defined in terms of scalar products of the i nput data in the original classification space. Thus, it is not necessary to specify $\Phi(.)$; instead of it kernel functions, $K(\boldsymbol{u} \cdot \boldsymbol{v})$, are used since they perform the transformation and scalar product in the transformed space in a single step.

There are several kernel functions that may be used in the definition of the new classification space as they are presented in table 1. Furthermore, Mercer theorem can be used to determine if a function could be used as kernel [19].

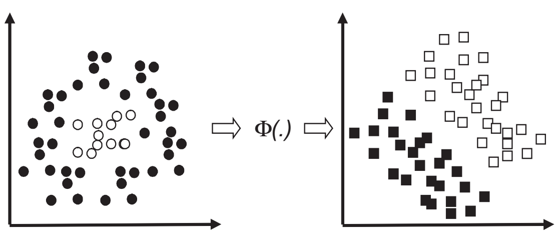

Figure 3 Data transformation in a new space where classes are linearly separable

Using an appropriate kernel function, SVM can separate data from different classes in this new space. Thus, linear classification algorithms can be extended to the non-linear cases by using an appropriate kernel function.

Table 1 Kernel functions

Name	*Function*
Polynomic kernel	$K(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u}\cdot\boldsymbol{v}+s)^d \quad s.t. \quad s>0$
Gaussian basis radial function (RBF)	$K(\boldsymbol{u},\boldsymbol{v}) = e^{-\lvert \boldsymbol{u}-\boldsymbol{v}\rvert^2/2\sigma^2}$
Sigmoid	$K(\boldsymbol{u},\boldsymbol{v}) = \tanh\,(\boldsymbol{u}\cdot\boldsymbol{v}+\Theta)$

When a RBF is chosen as kernel function (as in section five) two parameters (the error penalization constant C and the kernel parameter σ) have to be set to tune the classification algorithm. More extended information about SVM and kernel based methods is in [18, 19].

Features obtained from measurements

Since it is supposed that the only available measurements are voltages and currents at the substation, these have to be processed to obtain a set of useful descriptors. In the following subsections these features are described and the dependences with the fault resistance and distance are analyzed. For this purpose, faults in the main feeder of circuit of figure 4 have simulated under different conditions of resistance.

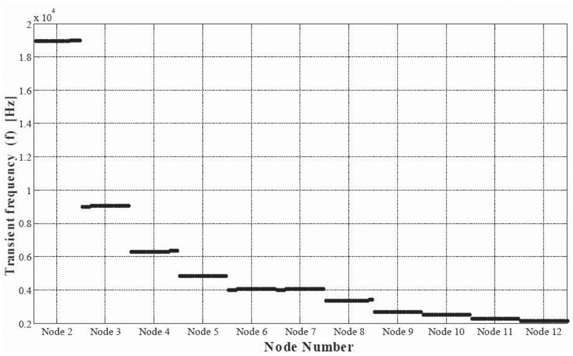

Figure 4 Transient frequency at the faulted phase Vs the distance (nodes along the main feeder) and fault resistance [0.5 Ω, 40 Ω]

Frequency of the transient

The natural frequency during the transient (f) can be associated to the distance to the fault, since it is directly related to the inductive and capacitive parameters of the remaining circuit in presence of a fault [20]. The frequency is obtained from the maximum energy detail of the decomposition of voltage transient using the Discrete Wavelet Transform (DWT) and the Fast Fourier Transform (FFT) [21, 22, 23]. In this paper, only the frequency has been used because its independence with the resistive parameters of the system, and consequently its relation with the distance. The non-dependence of frequency with the resistance can be observed in figure 4. Faults simulated at each of the 12 nodes along the main feeder of the power system in figure 15 and using 21 different fault resistance values from 0.5Ω t to 40Ω are presented.

Variation of rms values

The variations of the fundamental component of current (*ΔI*), voltage (*ΔV)* and apparent power (*ΔS*) have also been used as descriptors. These are defined from the subtraction of rms values during the fault and pre-fault steady states [4]. Because of the influence of the fault resistance the variation in the three phases has been used as descriptors instead of only the faulted ones.

Figures 5 and 6 shows a single phase faults on phase A, located at node 10 and 2, of the test power system represented in figure 13, respectively. These two figures illustrate two cases where the value of *ΔV* in the faulted phase is the same (-4488V) for different fault conditions (location and fault resistance). Nevertheless, the evolution of non-faulted phases is different. Variation of current (*ΔI*) and apparent power (*ΔS*) may experience similar effects.

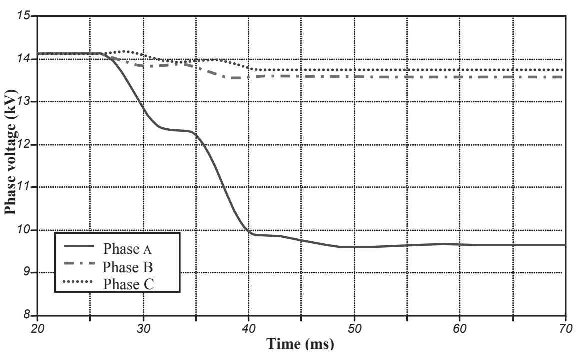

Figure 5 Single-phase fault in phase A at node 10. Fault resistance of 0,5 Ω

Figures 7, 8 and 9 represent the evolution of *ΔI*, *ΔV* and *ΔS* in case of single-phase faults along the 12 nodes main feeder for 21 values of fault resistance.

From figures 7, 8 and 9 it is noticed the multiple-estimation problem if only one of these attributes (*ΔV*, *ΔI*, *ΔS*) is used. Latter, in the application example, variations in the three phases are used to avoid the multiple estimation problem, by locating only one faulted zone.

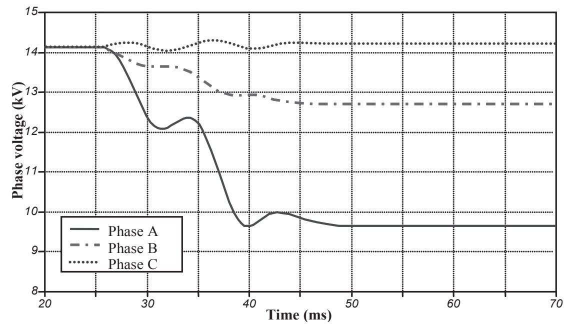

Figure 6 Single-phase fault in phase A at node 2. Fault resistance of 17.86 Ω

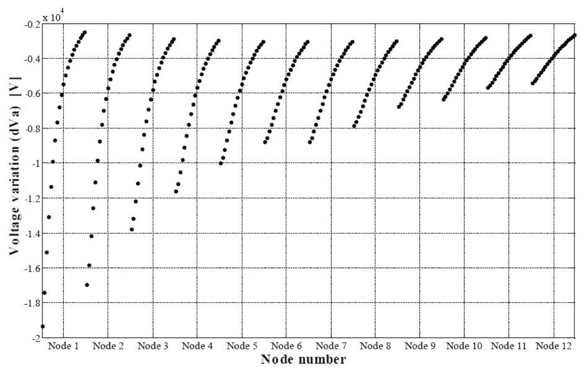

Figure 7 Voltage variation at the faulted phase Vs the distance (nodes along the main feeder) and fault resistance [0.5Ω, 40Ω]

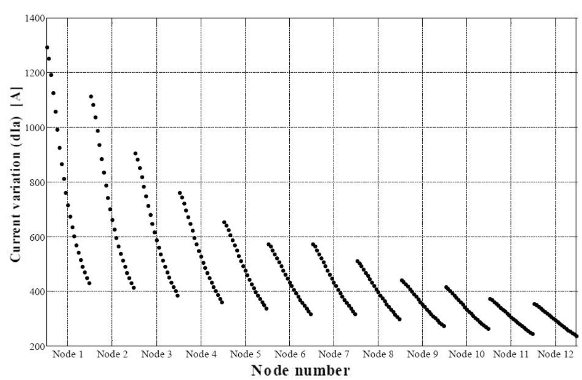

Figure 8 Current variation at the faulted phase Vs the distance (nodes along the main feeder) and fault resistance [0.5 Ω, 40 Ω]

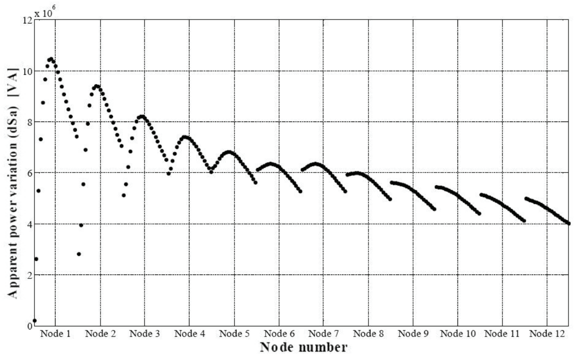

Figure 9 Apparent power variation at the faulted phase Vs the distance (nodes along the main feeder) and fault resistance [0.5 Ω, 40 Ω]

Fault reactance

The value of the reactance (*Xf*) seen from the substation during the fault is also related to the fault distance as presented in figure 10. It has been used as basis of the classical fault location algorithms [6].

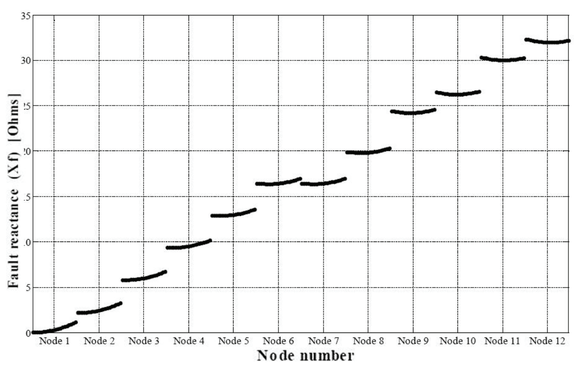

Figure 10 Fault reactance at the faulted phase Vs the distance (nodes along the main feeder) and fault resistance [0.5 Ω, 40 Ω]

In case of single phase to ground faults, reactance is computed by using phase values; otherwise, reactance is computed by using line-to-line magnitudes. The behavior of reactance and frequency versus the distance is quite similar.

Methodological approach for faulted zone location

The proposed strategy is based on the use of the available measurements of current and voltage at the power distribution substation. A selection of features extracted from these measures and sensible to fault location is proposed to train SVM based classifier to relate the faulted section with measures at the substation. Having this goal the following six steps are applied:

Step one: Zone definition in the distribution system

The power system is subdivided into significant zones to be identified in the fault location process. The zone definition criterion considers the following aspects: system topology, presence of protective devices, feeder lengths and other maintenance crew criteria as the maximum search length. It is recommended to subdivide laterals and long feeders into several zones according to the presence of protective devices and the representativeness of data available to train the SVM based classifier. In the presence of laterals it is also recommended to define a zone for each one if enough data is available to train the SVM to effectively avoid the multiple estimation problem.

Following the exposed, the system in the example has been divided in seven zones as presented in figure 13. If the faulted zone is correctly identified, the multiple estimation problem is avoided.

Step two: Structure of the SVM for fault location

Two different configurations of SVM for fault location are proposed. The first one has two serial stages to deal with different fault types (figure 11). The first stage recognizes the fault type, whereas the second one consists of four SVM specialized classifiers trained to locate the fault zone according: SVM PhG (Phase to ground classifier), SVM 2Ph (Phase to phase classifier), SVM 2PhG (Two phase to ground classifier) and SVM 3Ph (Three phase classifier).

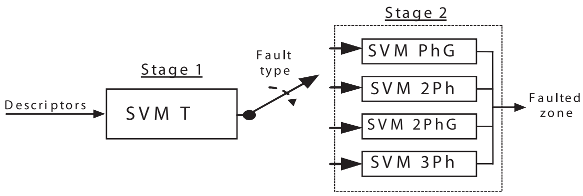

Figure 11 Fault type dependable SVM fault locator – DFL

The second structure proposed and tested is a simpler non-fault type dependable architecture based on the use of only one SVM to deal with all possible fault types (figure 12). Both fault locator configurations were tested in the example.

Figure 12 Non dependable SVM fault locator – NDFL

Step three: Training data

Measurements used to train the SVM are registered at the power substation but also simulations of faults (in this case a precise model is needed) can be used. In training stage, it is important to have registers from all the previous defined zones in order to train adequately the SVM based classifier to identify faults in any zone of the power distribution system.

Step four: Feature selection

All the descriptors or features previously presented are related to the fault location but according to system characteristics (load, heterogeneity, etc) the importance of them can vary. In section V the significance of them has been analyzed in the two different architectures for fault location in an application example.

Step five: Training

This step is performed by using a training set of disturbances characterized by descriptors ($\boldsymbol{x}_i$), and labels associated to the corresponding faulted zone ($\boldsymbol{y}_i$). This training set has to be complete and balanced. Complete in the sense that the training data set covers the whole search space and balanced to avoid a biased training of specific zones. Training is subdivided in two parts a) selection of the penalization parameter (C) and the kernel function (see equation 6 and table I) and b) the definition of the support vectors. These parts are described as follows:

a. *Selection of C and* σ **:** Grid search and cross validation have been used to select the best combination of C and the RBF kernel the parameter σ according to the method described in [19].

Grid search is defined as the variation of the two parameters (C and σ) in a feasible solution interval ($2^4<C<2^{32}$ and $2^{-6}<\sigma<2^6$). These two intervals originate a bidimensional grid of possible combinations of C and σ (the search space). When a maximum function performance is obtained, the C and σ values are selected as the best for such set of input descriptors. Cross validation is applied to obtain the performance for each pair (C, σ). This technique consists on a subdivision of the training set in n subsets. Training is performed by using (n-1) subsets. The remaining subset is used in the validation step. This procedure is repeated n times by using a different subset in the validation step and consequently a different combination of subsets in training. The performance is then obtained as the mean value of the obtained in the n tests.

Performance has been computed according to (8).

$$perfomance = \frac{Number\ of\ elements\ well\ classified}{Total\ number\ of\ elements} \tag{8}$$

b. *Definition of the support vectors:* Having obtained the best performance value, C and σ are selected to set the SVM. Next the SVM is trained by using all the training data. Train a SVM means to obtain the support vectors which helps to separate classes as presented in figure 1.

The previous training process has to be repeated for each one of the 31 subsets of descriptors considered in the application example.

Step six: Test the SVM for fault location

Having trained the SVM based fault locators presented in figures 11 and 12 by using the entire training set as presented in step five, an additional precision test has been carried by using new data not used in training test. The performance index is the same as the defined in equation (8).

Application example of the SVM based fault locator

Test system and dataset

A 25 kV power distribution system from Saskpower, Canada [5] (figure 13) has been used to test the fault location approach. The proposed system contains a three phase main feeder from node 1 to node 12, single-phase tapped loads at nodes 1, 2 and 8, and a three-phase taped load at node 12. It also has three laterals, a three-phase one in node 6 and two more single-phase at nodes 9 and 10. The length of the main feeder is approximately 37 km. A complete dataset have been obtained from simulation of faults in each node under different conditions. Single, double and three phase faults have been simulated by using 21 different values of fault resistances from 0,5 to 40 ohms [24]. Using Alternative Transients Program (ATP) [25] in an integrated environment linking ATP and Matlab, to automate the generation of faults [26], has simulated the power system. Table 2 summarizes the dataset used to train the SVM.

Table 2 Data sets used for training the svm fault locators

Fault type	*Fault locator*	*Zone number in the distribution system*							
		1	2	3	4	5	6	7	Total
P-G	SVM PhG	60	45	75	15	10	10	10	225
P-P	SVM 2Ph	60	45	75	-	-	-	-	180
2P-G	SVM 2PhG	60	45	75	-	-	-	-	180
3P	SVM 3Ph	20	15	25	-	-	-	-	60
All	SVM G	200	150	250	15	10	10	10	645

In case of zones from 4 to 7 there is not data of phase faults because these zones correspond to single-phase laterals

Definition of zones

The circuit has been subdivided in seven zones as it is depicted in figure 13. The main feeder was subdivided on three zones by using the criteria of reduce the possible location of the fault. Zone one comprises four nodes, zone two groups three nodes and zone three the last six nodes. These zones, in the main feeder, have been defined to overcome the multiple-estimation problem.

Laterals are also subdivided in zones according to the same criteria. Zones 5, 6 and 7 are defined to avoid multiple estimations of the fault locations caused by the radial topology of the system. In zone 4 the multiple estimation problem is not completely avoided due to the existence of sections with similar characteristics (from node 18 to node 19 and from node 18 to node 20). An additional zone has been defined with nodes 24 and 25 (Zone 7).

A constrain to the zone definition is the availability of data to train the SVM based locator. According to the exposed in section four, zone definition can also take into account the criteria of the operation staff and the maintenance crew.

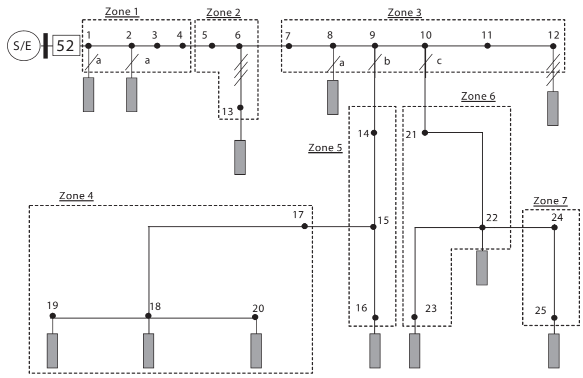

Figure 13 25 kV power distribution system from Saskpower, Canada

Training the SVM based fault locator

The performance of the zone fault locator has been analyzed by training and testing the SVM based classifier. Table 3 presents the data used in testing and training steps.

Table 3 Data sets used for training and testing the svm fault locator

Fault type	*Number of training elements*	*Number of testing elements*
Phase to ground (SVM PhG)	225	846
Phase to phase (SVM 2Ph)	180	639
Two phase to ground (SVM 2PhG)	180	639
Three phase (SVM 3Ph)	60	213
All fault types (SVM G)	645	2337

Stage one SVM based classifier in figure 11 has been trained to identify the fault type by using only two descriptors (ΔV and ΔI). Precision in testing step equals to the unity (no errors were obtained). On the other hand SVMs of stage 2 (figure 12) and the SVM G (see figure 12), have been trained by using data in table 2 for different combinations of descriptors in order to obtain the best combination. The training set was selected by using only five fault registers corresponding to fault resistance of 0.5, 5, 15, 25 and 40 ohms. The higher the number of training data, covering uniformly the whole set of zones, the better the precision of the precision test.

Result analysis

Table 4 summarizes the results. Groups of descriptors labeled as numbers 13, 19, and 26 presents the highest precision (perfect according to equation 14). All of these combination of descriptors which includes the reactance (Xf) appear as a good set of descriptors to be used to classification purposes. From table 4, in case of one the groups of descriptors (Group 10, composed by ΔV and ΔI), which has better performance, but not equal to unity, the desegregated test results are presented in table 5. These results correspond to the seven possible zones in case of phase to ground fault (SVM PhG fault locator).

From tables 2 and 5, it is noticed that having 225 sets of training elements and testing with 846 new sets, the average performance is around 99%. However, the classification performance in zones four and five are lower, but the precision is higher than 84 %. Table 6, presents the confusion matrix for phase to ground faults in table 5. The diagonal corresponds to the well-classified data, while the elements off the diagonal correspond to error in classification. It shows how data from zones four and five are not well classified using the variation of current (ΔI) and voltage (ΔV) as the only descriptors. From the analysis it is possible to determine that those wrong classified data in zone four, corresponds to faults in node 18 with low fault resistance value. Similarly, wrong classified data in zone five, corresponds to high resistance faults at node 16. However, it is noticed that by using a simple pre-processing of the voltage and current time signals to obtain descriptors as reactance (Xf), frequency (f) or variation of apparent power (ΔS) it is possible to obtain a perfect performance of the SVM fault locator under the presented conditions and to reduce the multiple-estimation problem.

Table 4 Extensive test of svm based fault locator

Group	Set of descriptors	SVM-PhG	SVM- 2Ph	SVM- 2PhG	SVM-3Ph	SVM-G	Group	Set of descriptors	SVM-PhG	SVM- 2Ph	SVM- 2PhG	SVM-3Ph	SVM-G
1	*ΔS*	0.924	0.914	1.000	0.906	0.940	17	*ΔS, ΔV, Xf*	1.000	1.000	1.000	1.000	0.999
2	*ΔV*	0.898	1.000	1.000	0.690	0.868	18	*ΔS, ΔV, f*	0.991	1.000	1.000	0.995	0.995
3	*ΔI*	0.879	1.000	0.995	1.000	0.869	19	*ΔS, ΔI, Xf*	1.000	1.000	1.000	1.000	1.000
4	*Xf*	1.000	1.000	1.000	1.000	0.965	20	*ΔS, ΔI, f*	0.978	1.000	1.000	1.000	0.995
5	*f*	0.981	1.000	1.000	1.000	0.981	21	*ΔS, Xf, f*	0.994	1.000	1.000	1.000	0.997
6	*ΔS, ΔV*	0.998	1.000	1.000	1.000	0.997	22	*ΔV, ΔI, Xf*	1.000	1.000	1.000	1.000	0.998
7	*ΔS, ΔI*	0.950	1.000	1.000	0.986	0.994	23	*ΔV, ΔI, f*	0.979	1.000	1.000	1.000	0.993
8	*ΔS, Xf*	1.000	0.992	1.000	1.000	0.998	24	*ΔV, Xf, f*	0.993	1.000	1.000	1.000	0.989
9	*ΔS, f*	0.980	1.000	0.998	0.986	0.983	25	*ΔI, Xf, f*	0.993	1.000	1.000	1.000	0.998
10	*ΔV, ΔI*	0.991	1.000	1.000	1.000	0.996	26	ΔS, ΔV, ΔI, Xf	1.000	1.000	1.000	1.000	1.000
11	*ΔV, Xf*	1.000	1.000	1.000	1.000	0.995	27	*ΔS, ΔV, ΔI, f*	0.996	1.000	1.000	1.000	0.991
12	*ΔV, f*	0.991	1.000	1.000	1.000	0.984	28	*ΔS, ΔV, Xf, f*	0.996	1.000	1.000	1.000	0.997
13	*ΔI, Xf*	1.000	1.000	1.000	1.000	1.000	29	*ΔS, ΔI, Xf, f*	0.996	1.000	1.000	1.000	1.000
14	*ΔI, f*	0.961	1.000	1.000	0.995	0.985	30	*ΔV, ΔI, Xf, f*	0.994	1.000	1.000	1.000	0.996
15	*Xf, f*	0.996	1.000	1.000	1.000	0.995	31	*ΔS, ΔV, ΔI, Xf, f*	0.983	1.000	1.000	1.000	0.995
16	*ΔS, ΔV, ΔI*	0.983	1.000	1.000	1.000	0.994							

Table 5 Desegregated test results in case of SVM PhG fault locator – Group of descriptors number 10

Faulted zone	*Test data*	*Well classified data*	*Precision in testing*
Zone 1	192	192	1
Zone 2	207	207	1
Zone 3	282	282	1
Zone 4	48	44	0.916
Zone 5	53	45	0.849
Zone 6	32	32	1
Zone 7	32	32	1
Total	846	838	0.991

Table 6 Confusion matrix of SVM PhG – Group of descriptors 10

		Zone number in the distribution system							***Total***
		1	2	3	4	5	6	7	
Zone number recognized by SVM 1	1	192	0	0	0	0	0	0	192
	2	0	207	0	0	0	0	0	207
	3	0	0	282	0	0	0	0	282
	4	0	0	0	44	8	0	0	52
	5	0	0	0	4	45	0	0	49
	6	0	0	0	0	0	32	0	32
	7	0	0	0	0	0	0	32	32
Total		192	207	282	48	53	32	32	846

Conclusions

This paper presents a methodological approach to locate the faulted zone based on the use of only the available measurements of current and voltage at the power substation. These signals are characterized to obtain descriptors as the variation of the effective values of voltage, current and apparent power, the reactance as seen from the substation in the steady state of fault and the frequency of the transient caused by the fault.

It is shown that the proposed approach has perfect performance in case of the analyzed example. Support Vector Machine based classifier has been tested to identify faulted zones and the obtained results show a good performance. The proposed approach allows reducing the multiple estimation problem of the fault location by using a low cost implementation strategy. This approach also contributes to improve the power continuity indexes in distribution systems by the opportune fault location.

References

1. M. Bollen. *Understanding Power Quality Problems: Voltages Sags and Interruptions*. IEEE Press.New York. 2000. pp. 5-16.
2. J. Janos. *Fault detection and diagnosis in Engineering Systems*. Marcel Dekker, Inc. New York. 1998. pp. 14-26.
3. Power System Relaying Committee. IEEE Std C37.114. *IEEE Guide for Determining Fault Location on AC Transmission and distribution Lines*. 2004. pp. 8-23.
4. J. Mora, G. Carrillo, B. Barrera. "Fault Location in Power distribution Systems Using a Learning Algorithm for Multivariable Data Analysis". *IEEE Trans. on Power Delivery*. Vol. 22. 2007. pp. 1715-1721.
5. R. Das. *Determining the Locations of Faults in distribution Systems*. Ph.D. Dissertation, University of Saskatchewan Saskatoon, Canada, Spring 1998. pp. 16-73
6. A. R. Warrington, C. Van. "*Protective relays. Their theory and practice"*. Vol I, Chapman and Hall Ltda. London. 1968. pp. 125-146.
7. K. Srinivasan, A. St-Jacques. "A new fault location algorithm for radial transmission lines with loads". *IEEE Trans. on Power Delivery*. Vol. 8.1989. pp. 1676-1682.
8. J. Zhu, D. Lubkeman, A. Girgis. "Automated fault location and diagnosis on electric power distribution feeders". *IEEE Trans. on Power Delivery*. 1997. pp. 801–809.
9. R. K Aggarwal, Y. Aslan, A.T. Johns. "An interactive approach to fault location on overhead distribution lines

with load taps". *IEE Developments in Power System Protection. Conference* Publication No. 434.1997. pp. 184-187.

10. M. Saha, E. Rosolowski. *Method and device of fault location for distribution networks. 2002*. US Patent number 6,483,435 B2.
11. M. Son Choi, S. Lee, D. Lee, B. Jin "A new fault location algorithm using ΔIrect circuit analysis for distribution systems". *IEEE Trans. on Power Systems*. 2004. pp. 35-41.
12. J. Mora, L. Pérez, G. Carrillo. "Fault location in power distribution systems using ANFIS nets and current patterns". *IEEE PES Transmission and distribution Conference and Exposition: Latin America. Caracas. 2006, ISBN 1-4244-0288-3.* pp. 1-6
13. P. Jarventausta, P. Verho. "Using fuzzy sets to model uncertainty in fault location process of distribution networks". *IEEE Trans. Power Delivery*. Vol. 9. 1994. pp. 954–960.
14. C. Chien, S. Chen, Y. Lin. "Using Bayesian network for fault location on distribution feeder" *IEEE Trans. on Power Delivery.* Vol. 17. 2002. pp. 785 – 793.
15. Y. Hsu, F. Liu, Y. Chien, J. Liu, J. Lin, H. Yu, R. Kuo. "An expert system for locating distribution system faults" *IEEE/PES Summer Meeting*. Vol. 1. 1990. pp. 366–372.
16. J. Peng, C. Chien, T. Tseng. "Rough set theory for data mining for fault diagnosis on distribution feeder" *IEE Generation, Transmission and Distribution*, - Vol. 151. 2004. pp. 689 – 697.
17. Y. Liu, N. Schulz. "Knowledge-based system for distribution system outage locating using comprehensive information". *IEEE Trans. on Power Systems*, Vol. 17. 2002. pp. 451–456.
18. C. Burges. *A tutorial on Support Vector Machines for Pattern Recognition*. http://www.kernel-machines.org/. pp. 25-58 Consultada el 15 de abril de 2007.
19. S. Bernhard, S. Alex. *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, Cambridge. 2002. pp. 24-69
20. A. Greewood. *Electrical Transients in Power Systems*. Renselarch Polytechnic Institute. 2nd ed.1991. pp. 56-79
21. S. Sitharama, E. Cho, V. Phoha. "*Foundations of Wavelet Networks and Applications*" Chapman and Hall/CRC, CRC Press LLC. 2002. pp. 24-35
22. A. Oppenheim, S. Willsky, S. Hamid Nawab "*Signals and Systems*". International Edition. Prentice Hall. 1997. pp. 45-101.
23. L. Sousa, J. Martins, V. Fernao. "A network distribution Power System Fault location based on Neural Eigenvalue Algorithm". *2003 IEEE Bologna Power Tech.* pp. 523-531
24. J. B. Dagenhart. "The 40- Ground-Fault Phenomenon" *IEEE Transactions on Industry Applications.* Vol. 36. 2000. pp 30-32.
25. H. W. Dommel. "*EMTP Theory Book*". 2nd ed. Vancouver. BC: Microtran Power System Analysis Corporation. 1992. pp. 25-102
26. J. Mora, J Meléndez, J. Bedoya. "Extensive Events Database Development using ATP and Matlab to Fault Location in Power distribution Systems" *IEEE PES Transmission and Distribution Conference and Exposition: Latin America. Caracas. 2006*.pp. 1-6.