Evolutionary Optimization of a One-Class Classification System for Faults Recognition in Smart Grids

Enrico De Santis[1], Gianluca Distante[1], Fabio Massimo Frattale Mascioli[1], Alireza Sadeghian[2] and Antonello Rizzi[1]

[1]*Dept. of Information Engineering, Electronics, and Telecommunications, "Sapienza" University of Rome, Via Eudossiana 18, 00184 Rome, Italy*

[2]*Dept. of Computer Science, Ryerson University, ON M5B 2K3, Toronto, Canada*

Keywords: Evolutionary Optimization, One Class Classification, Smart Grids, Fault Recognition.

Abstract: The Computational Intelligence paradigm has proven to be a useful approach when facing problems related to Smart Grids (SG). The modern SG systems are equipped with Smart Sensors scattered in the real-world power distribution lines that are able to take a fine-grained picture of the actual power grid state gathering a huge amount of heterogeneous data. Modeling and predicting general faults instances by means of processing structured patterns of faults data coming from Smart Sensors is a very challenging task. This paper deals with the problem of faults modeling and recognition on MV feeders in the real-world Smart Grid system that feeds the city of Rome, Italy. The faults recognition problem is faced by means of a One-Class classifier based on a modified *k*-means algorithm trained through an evolutive approach. Due to the nature of the specific data-driven problem at hand, a custom weighted dissimilarity measure designed to cope with mixed data type like numerical data, Time Series and categorical data is adopted. For the latter a Semantic Distance (SD) is proposed, capable to grasp semantical information from clustered data. A genetic algorithm is in charge to optimize system's performance. Tests were performed on data gathered over three years by ACEA Distribuzione S.p.A., the company that manages the power grid of Rome.

1 INTRODUCTION

The Smart Grid (SG) is one the best technological breakthrough concerning efficient and sustainable management of power grids. According to the definition of the Smart Grid European Technology Platform a SG should "*intelligently integrate the actions of all the connected users, generators, consumers and those that do both, in order to efficiently deliver sustainable economic and secure electricity supply*" (European Technology Plat., 2013). To reach that global goal the key word is the "*integration*" of technologies and research fields to add value to the power grid. The SG can be considered an evolution rather than a "revolution" (Energy Information Admin., 2013) with improvements in monitoring and control tasks, in communications, in optimization, in self-healing technologies and in the integration of the sustainable energy generation. This evolution process is possible if it will be reinforced by the symbiotic exchange with Information Communications Technologies (ICTs), that, with secure network technologies and powerful computer systems, will provide the "*nervous system*" and the "*brain*" of the actual power grid. Smart Sensors are the fundamental driving technology that together with wired and wireless network communications and cloud systems are able to take a fine grained picture not only of the power grid state but also of the surrounding environment. At this level of abstraction, the SG ecosystem act like a Complex System with an inherent non-linear and time-varying behavior emerging from heterogeneous elements with high degree of interaction, exchanging energy and information. Computational Intelligence (CI) techniques can face complex problems (Venayagamoorthy, 2011) and is a natural way to "inject" intelligence in artificial computing systems taking inspiration from the nature and providing capabilities like monitoring, control, decision making and adaptations (De Santis et al., 2013).

An important key issue in SGs is the Decision Support System (DSS), which is an expert system that provides decision support for the commanding

De Santis E., Distante G., Mascioli F., Sadeghian A. and Rizzi A..
Evolutionary Optimization of a One-Class Classification System for Faults Recognition in Smart Grids.
DOI: 10.5220/0005124800950103
In *Proceedings of the International Conference on Evolutionary Computation Theory and Applications* (ECTA-2014), pages 95-103
ISBN: 978-989-758-052-9

and dispatching system of the power grid. The information provided by the DSS can be used for Condition Based Maintenance (CBM) in the power grid (Raheja et al., 2006). Collecting heterogeneous measurements in modern SG systems is of paramount importance. As an instance, the available measurements can be used for dealing with various important pattern recognition and data mining problems on SGs, such as event classification (Afzal and Pothamsetty, 2012), or diagnostic systems for cables and accessories (Rizzi et al., 2009). On the basis of the specific type of considered data, different problem types could be formulated. In (Guikema et al., 2006) authors have established a relationship between environmental features and fault causes. A fault cause classifier based on the linear discriminant analysis (LDA) is proposed in (Cai and Chow, 2009). Information regarding weather conditions, longitude-latitude information, and measurements of physical quantities (e.g., currents and voltages) related to the power grid have been taken into account. The One-Class Quarter-Sphere SVM algorithm is proposed (Shahid et al., 2012) for faults classification in the power grid. The reported experimental evaluation is however performed on synthetically generated data only. This paper addresses this topic, facing the challenging problem of faults prediction and recognition on a real distribution network, in order to report in real time possible defects, before failures can occur, or as an off-line decision making aid, within the corporate strategic management procedures. The data set provided by Acea Distribuzione S.p.a (ACEA – the company managing the electrical network feeding the whole province of Rome, Italy) collects all the information considered by company's field experts as related to the events of a particular type of faults, namely Localized Faults (LF). This paper follows our previous work (De Santis et al., 2014) where the posed problem of faults recognition and prediction is framed as an unsupervised learning problem approached with the One Class Classification (OCC) paradigm (Khan and Madden, 2010) because of the availability only of positive or target instances (faults patterns). This modeling problem can be faced by synthesizing reasonable decision regions relying on a *k*-means clustering procedure in which the parameters of a suited dissimilarity measure and the boundaries of decision regions are optimized by a Genetic Algorithm, such that unseen target test patterns are recognized properly as faults or not. This paper focuses on two important issues: i) the initialization of *k*-means with an automatic procedure in order to find the optimal number *k* of clusters; ii) to find a more reliable dissimilarity measure for the categorical features of the faults patterns; to this aim, the Semantic Distance (SD) is adopted, addressing the problem of better grasping the semantic content of a well-formed cluster.

A brief review of the faults patterns is given in Sect. 2.1, while in Sec. 2 will be introduced the OCC system for fault recognition and the proposed initialization procedure for the *k*-means algorithm. In Sec. 2.4 is presented the weighted dissimilarity measure and the proposed SD for categorical features. In Sec. **??** it is shown and discussed the experimental results in terms of classifications performances comparing the well-known simple matching measure with the proposed semantic distance for categorical attributes. Finally, in the Sec. 4, conclusions are drawn.

2 THE ONE CLASS-CLASSIFICATION APPROACH FOR FAULTS DETECTION

2.1 The Fault Patterns

The ACEA power grid is constituted of backbones of uniform section exerting radially with the possibility of counter-supply if a branch is out of order. Each backbone of the power grid is supplied by two distinct Primary Stations (PS) and each half-line is protected against faults through the breakers. The underlined SG is equipped with Secondary Stations (SSs) located in the PSs able to collect faults data. A fault is related to the failure of the electrical insulation (e.g., cables insulation) that compromises the correct functioning of (part of) the grid. Therefore, a LF is actually a fault in which a physical element of the grid is permanently damaged causing long outages. LFs must be distinguished from both: i) "short outages" that are brief interruptions lasting more than one second and less than three minutes; ii) "transient outages" in which the interruptions don't exceed one second. The last ones can be caused, for example, by a transient fault of a cable's electrical insulation of very brief duration not causing a blackout.

The proposed one-class classifier is trained and tested on a dataset composed by 1180 LFs patterns structured in 20 different features. The features belong to different data types: categorical (nominal), quantitative (i.e., data belonging to a normed space) and times series (TSs). The last ones describes the sequence of short outages that are automatically registered by the protection systems as soon as they occur. LFs on MV feeders are characterized by heterogeneous data, including weather conditions, spatio-

Table 1: Considered features representing a FP.

Feature	Data type	Description
(1) Day start	Quantitative (Integer)	Day in which the LF was detected.
(2) Time start		Time stamp (minutes) in which the LF was detected.
(12) Current out of bounds		The maximum operating current of the backbone is less than or equal to 60% of the threshold "out of bounds", typically established at 90% of capacity.
(11) # Secondary Stations (SSs)		Number of out of service secondary stations due to the LF.
(3) Primary Station (PS) code	Categorical (String)	Unique backbone identifier.
(4) Protection tripped		Type of intervention of the protective device.
(5) Voltage line		Nominal voltage of the backbone.
(6) Type of element		Element that caused the damage.
(17) Cable section		Section of the cable, if applicable.
(7) Location element		Element positioning (aerial or underground).
(8) Material		Constituent material element (CU, AL).
(9) Primary station fault distance	Quantitative (Real)	Distance between the primary station and the geographical location of the LF.
(10) Median point		Fault location calculated as median point between two secondary stations
(13) Max. temperature		Maximum registered temperature.
(14) Min. temperature		Minimum registered temperature.
(15) Delta temperature		Difference between the maximum and minimum temperature.
(16) Rain		Millimeters of rainfall in a period of 24 hours preceding the LF.
(18) Backbone Electric Current		Extracted feature from Time Series of electric current values that flows in a given backbone of the considered power grid. It is the difference between the average of the current's value, in two consecutive temporal windows of twelve hours each one, before the fault.
(18) Interruptions (breaker)	TS (Integers sequence)	Sequence of opening events of the *breakers* in the primary station.
(19) Petersen alarms		Sequence of alarms detected by the device called "Petersen's coil" due to loss of electrical insulation on the power line.
(20) Saving interventions		Sequences of decisive interventions of the Petersen's coil which have prevented the LF.

temporal data (i.e., longitude-latitude pairs and time), physical data related to the state of power grid and its electric equipments (e.g., measured currents and voltages), and finally meteorological data. The whole database was provided by ACEA and contains data concerning a temporal period of three years across 2009–2011. This database was validated, by cleaning it from human errors and by completing in an appropriate way missing data. A detailed description of the considered features is provided in Table 1.

2.2 The OCC Classifier

The main idea in order to build a model of LF patterns in the considered SG is to use a clustering technique. In this work a modified version of k-means is proposed, capable to find a suitable partitions $P = \{C_1, C_2, \ldots, C_k\}$ of data set and to determine at the same time the optimal number of clusters k. The main assumption is that similar status of the SG have similar chances of generating a LF, reflecting the cluster model. The OCC System is designed to find a proper dicision region, namely the "faults space", $\mathcal{F}$, relying on the positions of target patterns denoting the LFs. A (one-class) classification problem instance is defined as a triple of disjoint sets, namely training set ($\mathcal{S}_{tr}$), validation set ($\mathcal{S}_{vs}$), and test set ($\mathcal{S}_{ts}$), all containing FP instances. Given a specific parameters setting, a classification model instance is synthesized on $\mathcal{S}_{tr}$ and it is validated on $\mathcal{S}_{vs}$. Finally, performance measures are computed on $\mathcal{S}_{ts}$. As depicted in the functional model (see Fig. 1) this paradigm is objectified by designing the OCC classifier as the composition of three modules wrapped in an optimization block. In order to synthetize the LF region, the learning procedure is leaded: 1) by the clustering module that operate an hard partition of $\mathcal{S}_{tr}$; 2) by the validation module operating on $\mathcal{S}_{vs}$, designed to refine the LF boundaries;

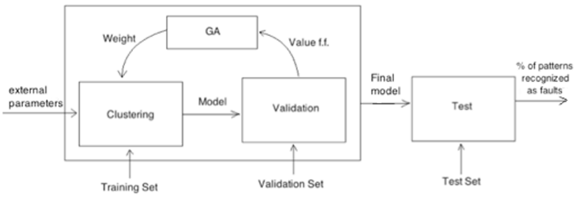


Figure 1: Block diagram depicting the optimized classification model synthesis.

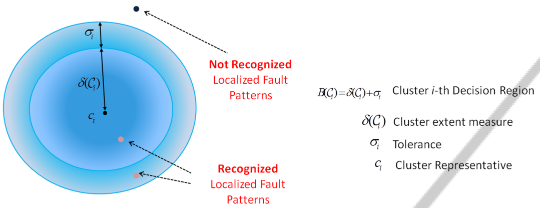


Figure 2: Cluster decision region and its characterizing parameters.

the decision rule (that leads the task of the *patterns assignment*) is based on the proximity of the LF pattern at hand to the clusters representative. Thus the core of the OCC system is the dissimilarity measure $d : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}^+$, reported in Sec. 2.4, that depends on a weighting parameter vector w. The decision regions $B(C_i)$ are derived from a "cluster extent" measure $\delta(C_i)$ characterizing the C_i clusters and summed to a tolerance parameter σ (thus $B(C_i)$=$\delta(C_i)$+ σ) that together to the dissimilarity weights belongs to the search space for the optimization algorithm. Here C_i is the i-th cluster ($i = 1, 2, \ldots, k$) and $\delta(C_i)$ is the average intra-cluster dissimilarity.

In this work in addition to the weights w and the σ parameters the search space is completed by a γ parameter controlling the proposed k-means initialization algorithm (see Sec. 2.3). Finally it is defined the search space, constituted by the all model's parameters, as $p = [w, \sigma, \gamma]$.

In this work the representative of the cluster, denoted as $c_i = R(C_i)$, is the MinSOD (Del Vescovo et al., 2014). So for each cluster C_i the representative one will be chosen as the pattern that belongs to the considered cluster and for which the sum of distances from the other patterns of the cluster has the lower value. A cluster representative c_i can be considered as a prototype of a *typical fault scenario* individuated in $\mathcal{S}_{tr}$. The decision rule to establish if a test pattern is a target pattern or not is performed computing its overall dissimilarity measure d from the representatives of all clusters C_i and verifying if it falls in the decision region (see Fig. 2) built up on the nearest cluster. A standard Genetic Algorithm is used in the learning phase in order to minimize the trade-off between the classification error rate on $\mathcal{S}_{vs}$ and the threshold σ'

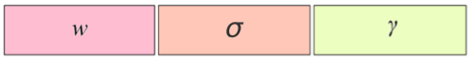


Figure 3: Composition of the chromosome.

value by means of the following Objective Function:

$$f(p) = \alpha\, ER(\mathcal{S}_{vs}) + (1-\alpha)\, \sigma', \qquad (1)$$

where $\alpha \in [0,1]$ is an external parameter controlling the importance in minimizing $ER(\mathcal{S}_{vs})$ versus σ'. In other words, α is a meta-parameter by which it is possible to control the relative importance in minimizing th error rate on $\mathcal{S}_{vs}$ or in minimizing the overall faults decision region extent. σ' is the threshold value normalized with respect to the diagonal D of the hypercube (see Sec. 2.4) of the overall space: $\sigma' = \sigma/D$. As concerns the chromosome coding (see Fig. 3), each individual of the population consists in the weights w_s, $s = (1, 2, \ldots, N_w)$ associated to each feature, where N_w is the number of the considered features, the value σ that is the threshold added at each "cluster extent" measure during the validation phase and the γ parameter mentioned above. The overall number of genes in an individual's chromosome is therefore l=$N_w + 2$. The functional dependencies between the discussed parameters, in the proposed OCC system, are:

$$\begin{cases} k_{opt} = k_{opt}(w, \gamma) \\ \sigma_i = \sigma_i(k_{opt}, w) \\ RR = RR(w, \sigma_i), \end{cases} \qquad (2)$$

where k_{opt} is the optimum number of clusters found by the k-means initialization algorithm described in the next Sec. and $RR = 1 - ER$ is the Recognition Rate of the proposed (one-class) classifier. The subscript index i covers the general case, not studied here, in which can be instantiated distinct thresholds values for different clusters.

2.3 The k-means Initialization Algorithm

It is well known that the k-means behaviour depends critically on both the number k of clusters, given as a fixed input, and on the position of the k initial clusters representatives. In the literature there are a wide range of algorithms for the initialization of the centroids of the k-means, each with its pros and cons (Dan Pelleg, 2000; Tibshirani et al., 2001; Laszlo and Mukherjee, 2006). The initialization criterion of centroids, here proposed, was initial inspired by (Barakbah and Kiyoki, 2009). The work is based on the idea to choose as centroids, the patterns that are furthest from each other. The provided version of the algorithm takes into account also the presences of *outliers*. To verify if the candidate centroid is an outlier we designed a simple decision rule defined by parameters: a, an integer value, and b, a real valued number ranging $[0,1]$. The parameter a indicates the minimum number of patterns that must enter in the circumference with center the pattern candidate as centroid and radius given by $b * d_{Pmax}$, where d_{Pmax} is the distance between the furthest pattern in the whole dataset. Hence if within the distance $b * d_{Pmax}$ there are more than a patterns then the candidate centroid is not an outlier. Other inputs to the overall algorithm are a scale parameter γ, the dissimilarity matrix **D** and the number of initially centroid k_{ini}. The algorithm tries to calculate the best positions of the centroids and their final number k_{opt} possibly decreasing the provided initial number (k_{ini}). The main steps are the following:

Algorithm input: The initial number of centroids k_{ini}, the dissimilarity matrix **D**, the $\gamma \in [0,1]$ parameter, a,b, d_{Pmax}.

Algorithm output: the k_{opt} centroids.

Choose a random pattern p_i among those available in $\mathcal{S}_{tr}$ and compute the pattern p_j furthest away from it.
while *centroid==not found* **do**
  if *p_j is not an outlier* **then**
    choose p_j as the first centroid;
    centroid=found;
  else
    choose as p_j the next pattern among those furthest away from p_i;
    centroid=not found;
  end
end
Choose as second centroid the pattern p_a furthest away from p_j.
while *centroid==not found* **do**
  if *p_a is not an outlier* **then**
    choose it as the as second centroid;
    centroid=found;
  else
    choose as p_a the next pattern furthest away from p_j;
    centroid=not found
  end
end
while $k < k_{ini}$ **do**
  choose as a possible centroid the pattern p_n whose sum of the distances to the other centroids, found earlier, is maximum;
  if *p_n is not an outlier* **then**
    choose it as the other centroid; $k = k+1$;
  else
    choose the next pattern whose sum of the distances to the other centroids found earlier, is maximum;
  end
end
Calculate $d_{Cmax} = d(p_j, p_a)$ as the distance between the first two centroids. Given the external parameter $\gamma \in [0,1]$
for *i=1; $i < k$; i++* **do**
  for *j=1; $j < k$; j++* **do**
    if $d(p_i, p_j) \leq \gamma * d_{Cmax}$ **then**
      delete randomly one of the two considered centroid, $k = k_{ini} - 1$;
    end
  end
end
return the $k_{opt} = k$ centroids.

The k-means with the proposed representatives initialization can be seen as an hybrid between a k-clustering and a free clustering algorithm where, once fixed an initial number of centroids, it returns an optimal number of centroids less or equal to the initial ones.

2.4 The Weighted Custom Dissimilarity Measure

The dissimilarity function between two patterns is of paramount importance in data driven modeling applications. Given two patterns x and y the wighted dissimilarity measure adopted in the proposed classifier is:

$$d(\underline{x},\underline{y};\underline{W}) = \sqrt{(\underline{x} \ominus \underline{y})\underline{WW}^T(\underline{x} \ominus \underline{y})^T} \tag{3}$$

$$d(\underline{x},\underline{y};\underline{W}) \in [0,D]$$

where $(x \ominus y)$ is a *Component-Wise* dissimilarity measure, i.e. a row vector containing the specific differences between homologues features. W is a diagonal square matrix of dimension $N_w \times N_w$, in which N_w is the number of weights. In Eq. 3 the maximum value for d is the diagonal of the hypercube, that is: $D = \sqrt{\sum_{i=1}^{N_w} w_i^2}$, where w_i are the features weights. The inner specific dissimilaritiy functions differ each other depending on the nature of each feature as explained in the following.

Quantitative (*real*). Given two normalized quantitative values v_i, v_j the distance between them is the absolute difference: $d_{i,j} = \mid v_i - v_j \mid$.

As regards the features "Day start" and "Time start" the distance is calculated through the circular difference. The value of these features is an integer number between 1 and 365 (*total days in one year*) for the former and between 1 and 1140 (*total minutes in one day*) for the latter. The circular distance between two numbers is defined as the minimum value between the calculated distance in a clockwise direction and the other calculated in counter clockwise.

Categorical (*nominal*). Categorical attributes, also referred to as nominal attributes, are attributes without a semantically valid ordering (see Tab. 1 for the

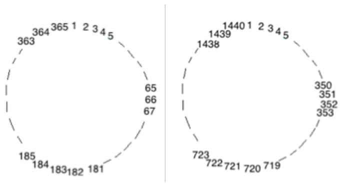


Figure 4: Sketch of circular domains for “Day start” and “Time start” features.

data treated as nominal). Let’s define c_i and c_j the values of the categorical feature for the patterns i-th and j-th, respectively. A one well-suited solution to compute a dissimilarity measure for categorical features is the Simple-Matching (SM) distance:

$$d_{i,j} = \begin{cases} 1 & if\ c_i \neq c_j \\ 0 & if\ c_i = c_j. \end{cases} \quad (4)$$

When measuring pattern-cluster dissimilarities (i. e. in the assignment of a pattern to a clusters) the Semantic Distance (SD) introduced in Sec. 2.4.1 is used.

Times Series TSs are characterized by a non-uniform sampling since they represent sequences of asynchronous events. As a consequence, usually they don’t share the same length. TSs are represented as real valued vectors containing the differences between short outages timestamps and the LF timestamp considered as a common reference. These values are normalized in the range $[0,1]$, dividing the values obtained by the total number of seconds in the temporal window considered. In order to measure the distance between two different TSs *(different in values and size)*, we use the Dynamic Time Warping (DTW) (Müller, 2007).

2.4.1 Semantic Distance for Categorical Data

The task of calculating a good similarity measure between categorical objects is challenging because of the difficulties to establish meaningful relations between them. The distance between two objects computed with the simple matching similarity measure (Eq. 4) is either 0 or 1. This often results in clusters with weak intra-similarity (Ng et al., 2007) and this may result in a loss of semantic content in a partition generated by a clustering algorithm. As concerns k-modes (Huang, 1998) algorithm, in the literature several frequency-based dissimilarity measures between categorical object are proposed (Cheng et al., 2004; Quang and Bao, 2004). The proposed dissimilarity measure for categorical objects is a frequency-based dissimilarity measure and follows the work (He et al., 2011) in which a features weighted k-modes algorithm is studied, where the weights are related to the frequency value of a category in a given cluster.

Let $N_{i,j}$ be the number of instances of the i-th value of the considered categorical feature F_{cc} in the cluster j-th (C_j) and let’s define $N_{max,j} = max(N_{1,j}, \ldots N_{n,j})$, where n is the number of the different values of F_{cc} present in C_j. We can finally define the SD between a categorical feature of the pattern P_h ($F_{cc}P_h$) and the cluster C_j as:

$$d_{F_{cc}P_h,C_j} = 1 - W_{i,j}, \ \ with \ d_{F_{cc}P_h,C_j} \in [0,1] \quad (5)$$

where $W_{i,j} = \frac{N_{i,j}}{N_{max,j}}$ is the fraction of values of the i-th category of the considered categorical feature in the j-th cluster with respect to the number of values of the most frequent category.
The SD take into account the statistical information of a given cluster and it is used like a pattern-cluster dissimilarity measure. Unlike the SM distance, the SD can span in the real valued range [0, 1]. Note that this distance is characterized by the statistical properties of the specific cluster under consideration. The SD can be intended as a “local metric”, since each cluster is characterized by its own statistic distribution of categorical values and thus it is characterized by its own weights that can change from one cluster to another.

For example, let us consider a categorical feature coding for one of four possible colors (red, green, blue or yellow) and let us consider the cluster depicted in Fig 5. By means of Eq. (5) it is possible

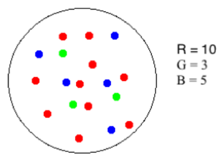


Figure 5: In this cluster the Yellow feature value is completely missing.

to compute the values of $W_{i,L}$ for each color (nominal attribute value), represented or not in the considered cluster and then the SD:

$$\begin{cases} W_{R,L} = \frac{10}{10}, W_{G,L} = \frac{3}{10}, W_{B,L} = \frac{5}{10} \\ if\ Color \ni C_L => W_{Color,L} = 0 \end{cases}$$

- if the value of $F_{cc}P_h$ is red: $d_{F_{cc}P_h,C_L} = 1 - \frac{10}{10} = 0$
- if the value of$F_{cc}P_h$ is green: $d_{F_{cc}P_h,C_L} = 1 - \frac{3}{10} = \frac{7}{10}$
- if the value of $F_{cc}P_h$ is blue: $d_{F_{cc}P_h,C_L} = 1 - \frac{5}{10} = \frac{1}{2}$
- if the value of $F_{cc}P_h$ is yellow: $d_{F_{cc}P_h,C_L} = 1 - \frac{0}{10} = 1$

3 EXPERIMENTAL RESULTS

3.1 Test on k-means Initialization Algorithm

The proposed initialization algorithm has been tested on a toy problem, where patterns are generated from three distinct Gaussian distributions, as depicted in Fig. 6. Setting the initial number of centroids K_{ini}=10, the proposed algorithm converges to an optimal number of clusters equal to 3 (see also Fig. 7).

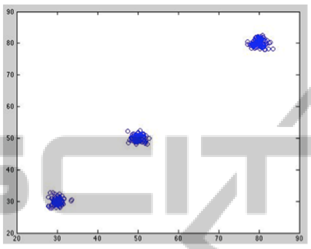
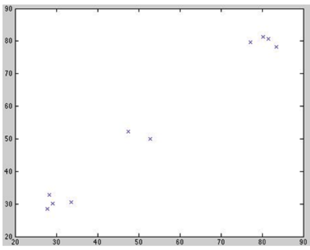
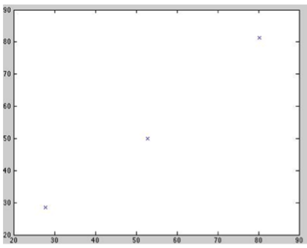


Figure 6: Patterns distribution in the considered toy problem.



(a) Centroids found before the close representatives removal step, with $k_{ini} = 10$.



(b) The final optimal centroids ($K_{opt} = 3$).

Figure 7

3.2 Tests on ACEA Dataset

In this section we report the first tests of the proposed OCC system on real data. The synthetized classification model should be able to correctly recognize fault patterns and, at the same time, to avoid raising wrong alarm signals, recognizing as faults system's measurements corresponding to normal operating conditions. Since non-faults patterns (negative instances) are not available in the ACEA dataset, in order to properly measure system performances, two distinct test sets have been employed. The first one, namely Ts1, is a subset of the available data set, i.e. a set of patterns labeled as faults. Obviously, the classification accuracy on this test set should be as high as possible. As concerns avoiding false positive misclassifications, a second test set, namely Ts2, has been created as a uniform random sampling of the whole input space domain, introducing constraints related to the physical network and environmental conditions in which the SG is located; thus Ts2 contains patterns labelled as both faults and non-faults. A high classification accuracy on Ts2 must be interpreted as a clue of a high number of false alarms, due to a too wide fault decision region. In close cooperation with the Acea experts, following their precious advice the LF model is trained on the features 1 to 4 and 6 to 18 (described in Tab. 1). Eighteen simulations, differing in the setting of α=*[0.3, 0.5, 0.7]* parameter and of $k_{ini} = [15, 10, 7]$ parameter, have been carried out, adopting both the proposed SD and the SM as pattern to clusters dissimilarity measure for categorical feature subspaces, yielding two different classification models, namely A and B, respectively. Best results are reported in Tab. 2. Although model B is characterized by a higher classification accuracy on faults patterns (Ts1), model A performs much better on Ts2 (the lower the better, when performance on Ts1 are comparable), since its fault decision region characterize much better faults pattern, with a much more limited extension in the whole input domain, avoiding thus to cover non-faults pattern. To confirm this interpretation we have computed the Davies-Bouldin index (Davies and Bouldin, 1979) on the training set partitions corresponding to models A and B. The Davies-Bouldin index (DB in Tab. 2) is a relative measure of compactness and separability of clusters (the lower the better). These measures confirm that clusters underlying the classification model A are more compact, yielding a much more effective and essential faults decision region. These results show that the proposed SD is much more suited in defining an effective inductive inference engine when dealing with categorical features subspaces, with respect to the plain SM distance.

4 CONCLUSIONS

In this paper we propose a MV lines faults recognition system as the core element of a Condition Based

Table 2: Result of the best simulation obtained with the SD and SM.

Model	α_f	k_i	k_{opt}	Categorical distance	γ	% Ts1	% Ts2	DB
A	0.3	7	7	Semantic Dissimilarity (SD)	0.4033	91.46%	23,12%	8.86
B	0.3	7	7	Simple Matching (SM)	0.4181	98,78%	35.76%	15.4

Maintenance procedure to be employed in the electric energy distribution network of Rome, Italy, managed by ACEA Distribuzione S.p.A. By relying on the OCC approach, the faults decision region is synthetized by partitioning the available samples of the training set. A suited pattern dissimilarity measure has been defined in order to deal with different features data types. The adopted clustering procedure is a modifed version of *k*-means, with a novel procedure for centroids initialization. A genetic algorithm is in charge to find the optimal value of the dissimilarity measure weights, as well as two parameters controlling the initial centroids positioning and the fault decision region extent, respectively. According to our tests, the new proposed method for *k*-means initialization shows a good reliability in finding automatically the best number of clusters and the best positions of the centroids. Furthermore, the proposed SD for categorical features subspaces performs better than the plain SM distance when used to define a pattern to cluster dissimilarity measure. Since faults decision region is synthetized starting from each cluster decision region, this measure has a key role in defining a proper inductive inference engine, and thus in improving the generalization capability of the recognition system. Future works will be focused on the definition of a suitable reliability classification measure, computed as the membership of incoming measures (patterns) to the fault decision region. Lastly, tests results performed on real data make us confident about further systems developments possibility, towards a final commissioning into the Rome electric energy distribution network.

ACKNOWLEDGEMENTS

The authors wish to thank Acea Distribuzione S.p.A. for providing the faults data and for their useful support during the OCC system design and test phases. Special thanks to Ing. Stefano Liotta, Chief Network Operation Division, and to Ing. Silvio Alessandroni, Chief Electric Power Distribution Remote Control Division.

REFERENCES

Afzal, M. and Pothamsetty, V. (2012). Analytics for distributed smart grid sensing. In *Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES*, pages 1–7.

Barakbah, A. and Kiyoki, Y. (2009). A pillar algorithm for k-means optimization by distance maximization for initial centroid designation. pages 61–68.

Cai, Y. and Chow, M.-Y. (2009). Exploratory analysis of massive data for distribution fault diagnosis in smart grids. In *Power Energy Society General Meeting, 2009. PES '09. IEEE*, pages 1–6.

Cheng, V., Li, C.-H., Kwok, J. T., and Li, C.-K. (2004). Dissimilarity learning for nominal data. *Pattern Recognition*, 37(7):1471 – 1477.

Dan Pelleg, A. M. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco. Morgan Kaufmann.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

De Santis, E., Rizzi, A., Livi, L., Sadeghian, A., and Frattale Mascioli, F. M. (2014). Fault recognition in smart grids by a one-class classification approach. In *2014 IEEE World Congress on Computational Intelligence*. IEEE.

De Santis, E., Rizzi, A., Sadeghian, A., and Frattale Mascioli, F. M. (2013). Genetic optimization of a fuzzy control system for energy flow management in micro-grids. In *2013 Joint IFSA World Congress and NAFIPS Annual Meeting*, pages 418–423. IEEE.

Del Vescovo, G., Livi, L., Frattale Mascioli, F. M., and Rizzi, A. (2014). On the Problem of Modeling Structured Data with the MinSOD Representative. *International Journal of Computer Theory and Engineering*, 6(1):9–14.

Energy Information Admin. (2013). *International Energy Outlook 2011 - Energy Information Administration, note=http://www.eia.gov/forecasts/ieo/index.cfm*.

European Technology Plat. (2013). *The SmartGrids European Technology Platform, note=http://www.smartgrids.eu/ETPSmartGrids*.

Guikema, S. D., Davidson, R. A., and Haibin, L. (2006). Statistical models of the effects of tree trimming on power system outages. *IEEE Transactions on Power Delivery*, 21(3):1549–1557.

He, Z., Xu, X., and Deng, S. (2011). Attribute value weighting in k-modes clustering. *Expert Systems with Applications*, 38(12):15365 – 15369.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values.

Khan, S. S. and Madden, M. G. (2010). A survey of recent trends in one class classification. In Coyle, L. and Freyne, J., editors, *Artificial Intelligence and Cognitive Science*, volume 6206 of *Lecture Notes in Computer Science*, pages 188–197. Springer Berlin Heidelberg.

Laszlo, M. and Mukherjee, S. (2006). A genetic algorithm using hyper-quadtrees for low-dimensional k-means clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):533–543.

Müller, M. (2007). Dynamic time warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer Berlin Heidelberg.

Ng, M. K., Junjie, M., Joshua, L., Huang, Z., and He, Z. (2007). On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Quang, L. and Bao, H. (2004). A conditional probability distribution-based dissimilarity measure for categorial data. In Dai, H., Srikant, R., and Zhang, C., editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*, pages 580–589. Springer Berlin Heidelberg.

Raheja, D., Llinas, J., Nagi, R., and Romanowski, C. (2006). Data fusion/data mining-based architecture for condition-based maintenance. *International Journal of Production Research*, 44(14):2869–2887.

Rizzi, A., Mascioli, F. M. F., Baldini, F., Mazzetti, C., and Bartnikas, R. (2009). Genetic optimization of a PD diagnostic system for cable accessories. *IEEE Transactions on Power Delivery*, 24(3):1728–1738.

Shahid, N., Aleem, S., Naqvi, I., and Zaffar, N. (2012). Support vector machine based fault detection amp; classification in smart grids. In *Globecom Workshops (GC Wkshps), 2012 IEEE*, pages 1526–1531.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Venayagamoorthy, G. K. (2011). Dynamic, stochastic, computational, and scalable technologies for smart grids. *IEEE Computational Intelligence Magazine*, 6(3):22–35.