Data-driven exploratory models of an electric distribution network for fault prediction and diagnosis

Noname manuscript No.
(will be inserted by the editor)

Data-driven exploratory models of an electric distribution network for fault prediction and diagnosis

Daniela Renga · Daniele Apiletti · Danilo Giordano · Matteo Nisi · Tao Huang · Yang Zhang · Marco Mellia · Elena Baralis

Received: date / Accepted: date

Abstract Data-driven models are becoming of fundamental importance in electric distribution networks to enable predictive maintenance, to perform effective diagnosis and to reduce related expenditures, with the final goal of improving the electric service efficiency and reliability to the benefit of both the citizens and the grid operators themselves.

This paper considers a dataset collected over 6 years in a real-world medium-voltage distribution network by the Supervisory Control And Data Acquisition (SCADA) system. A transparent, exploratory, and exhaustive data-mining workflow, based on data characterisation, time-windowing, association rule mining, and associative classification is proposed and experimentally evaluated to automatically identify correlations and build a prognostic-diagnostic model from the SCADA events occurring before and after specific service interruptions, i.e., network faults.

Our results, evaluated by both data-driven quality metrics and domain expert interpretations, highlight the capability to assess the limited predictive capability of the SCADA events for medium-voltage distribution networks, while their effective exploitation for diagnostic purposes is promising.

1 Introduction

Electric grid operators welcome predictive maintenance to avoid the costs of scheduled inspections and reactive maintenance interventions. To this aim, datasets describing the electric grid operations and historical data about failures and alarm signals are exploited to design predictive maintenance solutions. Although this data has been collected for different purposes, companies are interested in determining their predictive maintenance capability: to reduce management costs, to speed up intervention-time, and to improve efficiency

D. Renga, D. Apiletti D. Giordano T. Huang, Y. Zhang, M. Mellia, and E. Baralis are with the Politenico di Torino, Turin, Italy (email: name.surname@polito.it) · M. Nisi was with the Politecnico di Torino, Turin, Italy (e-mail: m.nisi@studenti.polito.it).

and reliability.
For our study, we rely on a Big-Data collection spanning over 6 years, collected by a leading Italian electric grid operator. The dataset describes the operations of the medium-voltage distribution network in northeastern Italy, and it records events and failure through the Supervisory Control And Data Acquisition (SCADA) system. Our aim is to assess whether this dataset could be exploited to (i) predict future electric network failures (predictive maintenance) and/or (ii) effectively diagnose the failures after it is reported by the maintenance system. Since the predictive capability of such dataset and the capability to model system degradation are unknown, we address these tasks by means of an exploratory and exhaustive predictive-maintenance analysis.
The main contributions of our study can be summarised as follows.

- We combine different exploratory approaches to evaluate the dataset capability of predicting the possible occurrence of future faults, the affected component, and the fault cause, and to perform such diagnosis after faults are detected.
- We exploit a transparent and exhaustive method based on *association rule mining* to automatically extract from the dataset all correlations, above specific statistical thresholds, among SCADA events occurring before each fault of interest (prognostic), and separately, after the faults (diagnostic).
- We design a transparent associative-classification model based on human-readable rules to predict and diagnose the type of fault and the affected component.
- We perform a thorough comparison between the prognostic and diagnostic approaches by means of both data-driven quality metrics and domain-expert analysis.

To the best of our knowledge, our work is the first study that investigates both the prognostic and diagnostic capabilities of a real-world large historical dataset collected by a Supervisory Control and Data Acquisition (SCADA) system in an electric grid, with respect to the occurrence of severe service interruptions. Thanks to the application of an exhaustive analysis methodology, i.e. the extraction of association rules among faults and events, we address the issue of providing smart grid operators an assessment of the exploitation potential of currently available datasets for predictive maintenance and diagnosis. The proposed methodology can be applied to similar datasets from any grid operator.

2 Related work

With the shift from the traditional electric grid to the Smart Grid paradigm where the energy generation becomes distributed and more complex consumers arise, huge efforts are required to deploy solutions to make the electric system more reliable and robust towards faults [1]. The remarkable advances in instrumentation, communication and data analysis have made it possible to introduce effective solutions for grid monitoring and management. Recently, several studies show the key role that data analytics and related applications

are acquiring in power networks [2,3]. However, few research efforts have been specifically devoted to predictive maintenance, whereas various studies are available about fault monitoring and characterisation. Some studies aim at performing fault detection in power networks based on historical weather data mining [4], on extreme learning machine models [5], or on electrical feature extraction techniques [6]. Authors in [7] deploy an effective method to detect faults in smart grids, trading off the need for reducing the volume of collected data, related to the Phasor measurement unit, and the need for keeping critical information. Other studies aim not only to detect faults, but also to further characterise them by identifying and exploiting significant features. Classifiers based on clustering and dissimilarity learning techniques [8] or on feature extraction algorithms [9] are used to analyse massive data to perform fault recognition or distribution fault diagnosis. In [10], authors apply a learning intelligent system for classification and characterisation of localised faults in Smart Grids. In addition, SCADA event datasets are typically considered and investigated for monitoring purposes and fault diagnosis in different types of systems, especially in renewable power plants [11–14].

Regarding failure prevention, several authors employed various data mining methods for fault prediction in the most disparate fields, from aircraft security to industrial systems to renewable power plants. Studies from the literature show that neural networks,support vector regression, generic data-driven based approaches, and techniques exploiting classifiers or association rule extraction can be adopted in fault prediction [11, 14–19]. In particular, techniques based on association rules learning are currently often applied in fault prediction of industrial systems, but no studies are available for electric MV distribution networks [18, 19]. A data mining method based on similarity association rule is proposed in [19] to detect sensor faults in power plants. In [18], authors propose a novel algorithm capable to deal with huge and sparse datasets, exploiting multidimensional time-series data to extract association rules without the need of reducing the data, hence possibly causing the loss of information and missing relevant rules. Nevertheless, the deployment of fault detection methods with prognostic purposes in MV distribution networks is not well investigated in the literature [1]. Only authors in [20] aim at reducing the outages in Medium Voltage distribution networks by exploiting rule-based, data mining and clustering techniques to design a method providing diagnostic and prognostic functions for Distribution Automation systems.

The first part of our study exploits a well known approach based on association rule mining. However, unlike the majority of the papers available in the literature, our work considers a MV distribution network aiming at jointly investigating both the diagnostic and prognostic potential with respect to electric service interruptions -in terms of fault type and affected component- of an already existing real-world historical dataset that collects events registered by the SCADA system. Furthermore, to the best of our knowledge, no work is available exploiting association rule classifiers to detect or predict faults in an electric distribution grid.

Fig. 1: Main steps of the proposed approach.

3 The proposed approach

Medium-voltage distribution networks provide extensive collections of events through their SCADA systems. Given such a dataset from a real-world scenario, a classic data mining task entails extracting useful patterns and correlations to improve the network maintenance. Hence the problem is twofold: (i) identifying a methodology able to automatically assess the prognostic and diagnostic potential of the provided dataset, and (ii) addressing such a task with transparent approaches, able to provide human-readable explanations of the identified patterns. Here, we preliminarly perform a *dataset characterisation* stage (Sec. 4), to identify the most frequent SCADA events, and how long we should monitor the events before and after a fault to collect useful information for the prognostic and diagnostic analysis. We present the main steps of our methodology in Figure 1: (i) a *time-based windowing* stage, where we process the raw SCADA dataset to extract time windows in the pre-fault and after-fault periods (Sec. 5), based on the evidence provided by the dataset characterisation stage. (ii) The *rule-mining extraction* process (Sec. 6) to identify in an automatic, exploratory, transparent and exhaustive way all the statistically relevant patterns and correlations representing the pre- and after-fault windows. Finally, (iii) an *associative-classifier-based analysis* (Sec. 7) is used to test the prognostic potential of the dataset.

4 Dataset

The dataset under analysis contains events recorded by the SCADA system of a leading Italian grid operator on its medium-voltage distribution network. The dataset spans over a period of 6 years (2010-2016), covering two northeastern Italian regions (Veneto and Friuli-Venezia-Giulia). The dataset contains by 3,901 faults of interest, 30 different affected components, 153,094 general SCADA events recorded during network operations. The SCADA events are divided into 67 different event types, with the generic failure event accounting for 79,833 events. Here we are interested in those faults: (i) lasting more than 180 seconds, (ii) with the location in the network identified, and (iii) with the cause determined. These events are named Permanent Service Interruptions (PSIs), tagged with a cause among 45 different reasons and linked to one among the 30 affected components.

We briefly characterise the dataset by analysing the distribution of PSIs causes and types of SCADA events. Considering the probability distribution of the most frequent causes of PSIs among the 45 available, we observe an heavy tailed shape, with the top 4 causes accounting for 75% of the PSIs, with “elec-

tric fault" being the most frequent cause (45%). More than 20% of PSIs are due to natural causes, such as weather issues, plant falls, snow overload, wind, and animal contacts. All these causes are unpredictable without contextual knowledge outside the electrical grid operational events. Furthermore, another 20% of PSIs are due to unknown "other causes" (second most frequent value). Regarding the probability distribution of the most common SCADA events types, it is similarly skewed, with about 75% of SCADA events belonging to just 6 different types, and with the most frequent one with a frequency above 30%.

5 Prognostic-diagnostic approach

Since we aim at investigating both the prognostic and diagnostic potential of SCADA events with respect to PSIs, we focus on the analysis of those events occurring both before and after a PSI, in the same portion of the network, under the assumption that the time and space correlations might capture causalities of the system.
In the time dimension, we define a time window preceding the occurrence of a PSI, denoted as *Pre-Fault Window* (PFW), and a time window immediately following the PSI, denoted as *After-Fault Window* (AFW). In the space dimension, we consider only SCADA events observed in the same portion of the network where the PSI occurs, i.e., reported by the same feeder as origin of the collected data, since according to the domain experts they are more likely to be correlated to the considered PSI. Considering that the grid operator is interested in predicting future PSIs occurring within the next month at most,we consider time windows of 1, 7, 30 days for PFW, and 1 hour, 1 day or 7 days for AFW. These values result from wider preliminary analyses, with the aim of capturing behaviours of the distribution network at different time scales of interest for domain experts of the electric grid company.
The time-window-based characterization of the SCADA event distribution during the PFWs and the AFWs shows that 60% of the PSIs have no SCADA events in their 7-day PFW, whereas 10% of the PSIs have no SCADA events in their 1-day AFW [21]. In addition, most diagnostic events occur in the 1-hour AFW. Finally, many events occur more then 1 week before the PSI (PFW); however, they include events generated as a consequence of other minor faults, i.e., they are in the AFW of non-permanent Service Interruptions. This happens in 60% of the cases for a 30-day PFW, and in 26% of cases for a 7-day PFW. These results suggests a limited *prognostic* potential of the SCADA events with respect to PSIs due to the few events, mostly time-unrelated. Conversely, the *diagnostic* exploitation seems better supported by more data that is recorded after the event of interest.

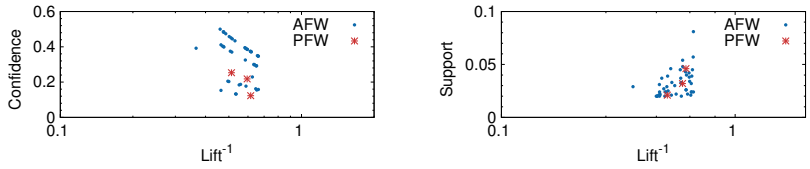
6 Rule Mining

To address challenges identified in Section 5, we exploited a transparent, exhaustive and exploratory data mining approach: association rule mining. In

the following, we describe the technique and evaluation metrics, as required by the scope of the current work, are defined as follows.

6.1 Association Rule Extraction

Let $\mathcal{D}$ be a dataset whose generic record r consists of a set of co-occurring events, i.e., events that occur in the same time window. Each event, also called *item*, is a couple *(attribute, value)*. In the current work, the *attribute* is either a SCADA event type, or an alleged cause, or a failed component, and the *value* is 1 if that attribute is true in the time window under exam (e.g., the SCADA event is present, the component failed, or the specific cause was determined), or 0 otherwise. Note that a SCADA event might represent another PSI or a minor fault occurring before or after the analysed PSI. An *itemset* I is a set of co-occurring events, failed components, and alleged causes among the records r in the dataset $\mathcal{D}$. Such set of items I in a PFW or, separately, in an AFW constitutes the input feature vector of the rule mining extraction.
The *support count* of an itemset I is the number of records r containing I. The *support* $s(I)$ of an itemset I is the percentage of records r containing I with respect to the total number of records r in the full dataset $\mathcal{D}$. An itemset is *frequent* when its support is greater than or equal to a minimum support threshold $MinSup$.
Association rule mining aims at identifying collections of itemsets (i.e., sets of co-occurring events) that are frequently present in the dataset under analysis, according to statistically relevant metrics. The extracted rules are all and only those adhering to the thresholds of statistical relevance defined as parameters of the mining process, hence being an exhaustive, thus powerful, exploratory approach within the boundaries of the problem formulation (i.e., itemset definition and threshold settings).
Association rules are usually represented in the form $X \rightarrow Y$, where X (rule antecedent) and Y (rule consequent) are disjoint itemsets (i.e., they include different attributes). To identify the most meaningful rules among those extracted by the mining process, quality measures can be exploited as ranking criteria. The following popular quality measures are used in the current work: rule support, confidence, and lift. *Rule support* $s(X, Y)$ is the percentage of records containing both X and Y. It represents the prior probability of $X \cup Y$, i.e., the support of the corresponding itemset $I = X \cup Y$ in the dataset. *Rule confidence* is the conditional probability of finding Y given X. It describes the strength of the implication and is given by $\mathrm{c}(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$ [22].
All and only association rules with support and confidence above (or equal to) a support threshold $MinSup$ and a confidence threshold $MinConf$ are to be extracted. Among those surviving the thresholds, a rank based on descending support, confidence and lift values can drive the attention to focus on the most statistically-relevant patterns. The *lift* [22] of a rule $X \rightarrow Y$ measures the (symmetric) correlation between antecedent and consequent. If lift(X,Y)=1, itemsets X and Y are not correlated, i.e., they are statistically independent. Lift values below 1 show a negative correlation between itemsets X and Y,



(a) Confidence vs lift^{-1} (b) Support vs lift^{-1}

Fig. 2: Association rules extracted from the 7-day PFW (a-b) and from the 1-day AFW (c-d), with causes or components as conclusion (x-axis in log scale).

while values above 1 indicate a positive correlation, with higher lift indicating stronger rules, hence typically more meaningful and interesting correlations.

6.2 Rule quality analysis

The analysis of the extracted rules has been performed for various parameter values. Due to space constraints, we report only the most meaningful results based on the rules obtained by (i) setting $MinSup$ 0.02, then focusing on rules (ii) whose lift is higher than 1.5, and (iii) having a cause or component as conclusion. We report the number of rules resulting from such selection in Figures 2a-2b for a 7-day PFW (red stars) and, for comparison, for an AFW of 1 day (blue dots). The scatter-plot shows support and confidence versus lift values. The *diagnostic* potential (AFW) is confirmed by a larger number of rules with better quality metrics with respect to the *prognostic* capability (PFW): overall, we have 45 rules in the AFW vs 3 in the PFW, with 50% max rule confidence in AFW vs 25% in PFW, 2.73 max lift value in AFW vs 1.9 in PFW, and 8% max support in AFW vs 4.5% in PFW. We inspect top rules according to lift, confidence and support with the help of domain experts from the grid company, allowing to transparently evaluate the correlation model and the prognostic-diagnostic approach.

6.3 Domain-expert analysis

From the results of Table 1, we can see that the most interesting combinations of the SCADA events and the components are related to PSI affecting the highest percentage of the components of the distribution network, i.e. electric conductors, including the aerial lines and cables.

Take the results of the opening of MV line for max current 2^{nd} threshold as example. It signifies that the system has a fault happening with a large current while the relay does not immediately trip the line. The fault current harms the

Table 1: Sample of association rules for PFW, considering causes of faults as consequent.

Antecedent	**Consequent**	**Supp** [%]	**Conf** [%]	**Lift**
Opening of MV line for max current 2^{nd} threshold	Conductor	2.05	25.32	1.95
RG CTO intervention	Conductor	3.23	21.80	1.68
Telesegn Disp 180	Cable	4.56	12.33	1.62

insulation of the equipment and over time, with repeated faults, the equipment will eventually introduce ground fault at any time in the future. The RG CTO intervention would have a similar effect with a higher magnitude, as the fault current in this case would be much larger.

As the antecedent events in Table 1 appear in the PFWs as the prognostic signals, it might be inferred that, after an introduction of fault in the system, the accumulated effects of the deterioration of the insulation over time would account for a noticeable amount of PSIs in the future. However, due to the lack of the time labels of the events, we cannot exclude other possible causes where that the PSI in study is a consequence of an immediately preceding fault related to another PSI.

By contrast, for the AFW, if the opening of MV line for max current 2^{nd} threshold has been recorded by the SCADA system, it is not surprising that the highest percentage of the related PSIs will affect the largest number of equipment, i.e. conductors in the system.

If we compare the last two items of the results in Table 2, we can clearly see that the confidence values are quite high, i.e., higher than 41%, which shows that the T-junction is the most vulnerable elements against the premises. However, if we check the SCADA event closely, the most decisive one for such a conclusion is the Permanent opening round 2^{nd} threshold, which indicates that the fault could not be cleared by itself and developed into a permanent one. For other types of fault, such as branch touching with the conductor, they can be cleared by themselves, and the automatic reclosing relay (RDA) would reclose the line and restore its services. On the other hand, the most typical permanent fault would be the fault on the equipment itself, in this case, the T-junction. Therefore, our study clearly captured the phenomenon in which the T-junction has a failure and generates fault current. Since it cannot be cleared by itself, the concerning relay permanently trips the relevant lines.

7 Associative classification analysis

Association rules extracted from the general-purpose prognostic-diagnostic approach, as in Sec. 6, describe all the correlations among attribute values above given thresholds (support and confidence) in the overall SCADA-event dataset. Those attributes include (i) the component affected by a fault and (ii) the reason of a fault, which are two desired target variables in our prognostic-diagnostic context.
Indeed, selecting only the subset of association rules whose heads are restricted

Table 2: Sample of association rules for AFW, considering causes of faults as consequent

Antecedent	**Consequent**	**Supp** [%]	**Conf** [%]	**Lift**
Opening of MV line for max current 2^{nd} threshold	Conductor	2.91	39.24	2.73
Opening of MV line for ground fault 2^{nd} threshold, Permanent opening ground 2^{nd} threshold, SC TC ground 2^{nd} threshold	T-junction	2.03	50	2.18
Permanent opening ground 2^{nd} threshold, SC TC ground 2^{nd} threshold	T-junction, electric fault	2.01	41.05	2.16

to the target attributes (i.e., Class Association Rules) can help [23]. However, for association rules, the target of mining is not predetermined. Instead, we introduce in this section an analysis based on the associative classification approach, which optimises the choice of the extracted rules with respect to a pre-determined target attribute, i.e., the class.
Such approach, in a predictive-maintenance context, specifically addresses the common question of automatically assessing the capability of a given dataset to be able to determine the component affected by a fault, or the specific reason for such a fault. This information is useful to reduce maintenance operations of the electric grid, for instance by providing the intervention team with suitable tools to fix the fault. The associative classifier exploited in our study is L3 [24]. According to this approach, association rules are exploited to build a classifier. While the rule consequent is restricted to the class labels, the rule antecedent represents a set of feature-value couples that should be matched to predict the considered class. We recall that features are represented by the presence or absence of each SCADA event, as registered in the time window under analysis, with a record for each fault. Labels can be (a) the type of component interested by the PSI, or (b) the PSI cause.

7.1 Data preparation

To apply the associative classifier, we perform a data-preparation workflow on the input dataset, that includes two steps: (i) class removal and (ii) feature selection.
Since the associative classifier is built upon the frequency of the events, many samples describing the behaviour of each class are required to effectively learn to predict a class label from the data. Hence, we remove classes with a number of samples less than a threshold S_n from the dataset. Avoiding such removal would lead to rules with limited interest, since they would model spurious behaviours, hardly capturing general patterns applicable to new data. Such pruning reduced the number of classes, from 30 components and 45 fault causes to the final numbers reported in Tab. 3, whose results have been obtained setting $S_n = 100$.
We next run a random-forest-based [25] feature selection provided in Scikit-Learn [26] to select the most relevant attributes, hence reducing the model complexity. We remove those features with an importance score lower than a threshold F_i with respect to the value of the most important feature. Tab. 3 reports the number of features available after the feature selection procedure with $F_i = 10\%$. A qualitative analysis of the selected attributes let us note that the SCADA event types describing the presence of other faults within the time window are included as valuable features for the PFW, while they are filtered out for the AFW. This is due to the window length, which is 1 week long for the PFW but only 1 hour long in the AFW.

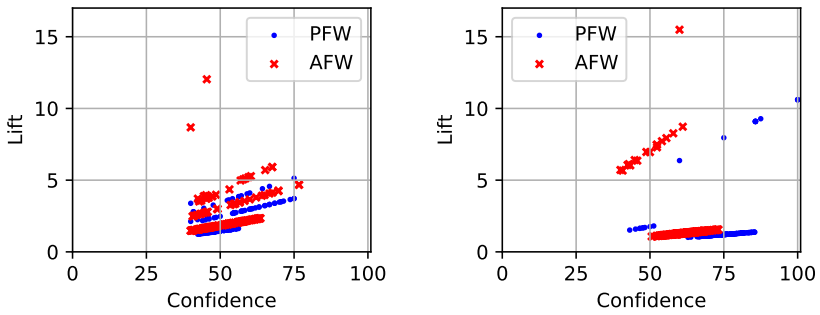
7.2 Experimental results

The associative classifier results have been analysed for different values of the minimum support and confidence thresholds. Specifically, we determined 5%

Table 3: Dataset description and number of rules selected by the associative classifier with $S_n = 100$, $F_i = 10\%$, $minsup = 5\%$, and $minconf = 40\%$.

Class label	Window	Attributes	Classes	Rules
Component	PFW	13	5	107
Component	AFW	19	10	200
Cause	PFW	17	3	195
Cause	AFW	20	9	291

and 40% to be, respectively, the highest values yielding to a non-empty set of rules in the first level of L3 [24]. Note that the minimum relative support threshold for the associative classifier is referred to each class, whereas for association rule mining in Sec. 6 the minimum support threshold was on the whole dataset. Hence, a 5% threshold for the associative classifier means that the corresponding rule will be extracted for a given class label only if its support is at least the 5% of the number of samples of that class label.
To evaluate the prognostic-diagnostic potential, we analyse the distribution of the quality metrics of the rules of the associative classifier, separately for PFW (prognosis) and AFW (diagnosis), and for each class label (faulty component or cause of fault). We report results in Fig. 3 by exploiting two quality metrics: confidence and lift, including also the support metric did not lead to additional insights. A common pattern in most cases, emerging from the comparison of the PFW (prognosis) and AFW (diagnosis) results, is that the diagnostic rules have better (higher) quality metrics. Specifically, when the target class is the component (see Fig. 3a), the trend is present not only in the reported confidence-lift plot, but also in any pair of metrics. The group of top quality rules always belong to the AFW, with much higher support and lift, and a slightly higher confidence. When the classifier targets the cause (see Fig. 3b), instead, AFW and PFW rules seems to have similar values for quality metrics. However, we should consider that the Cause-PFW classification problem is much easier due to the lower number of classes (3) with respect to



(a) Confidence vs Lift – Component (b) Confidence vs Lift – Cause

Fig. 3: Analysis of the quality metrics of the rules selected by the associative classifier.

the Cause-AFW (9 classes), as reported in Tab. 3. As a consequence, we can expect PFW rules to be generally less powerful, and the dataset predictive potential to be limited with respect to its diagnostic capability.
To this aim, we train an L3-based prediction model on 70% of the dataset, and test the model-generated rules on the remaining 30% to predict either components or causes of faults. Classification results confirm the poor predictive power indicated by the rule-mining analysis. The cause-prediction L3 model provides a label for 927 samples (96% of the test set) with an average precision of 0.34 and an average recall of 0.54, with an unbalanced prediction towards the majority class. The component-prediction L3 model yielded even lower results, with an average precision as low as 0.01, an average recall of 0.12, and an extremely unbalanced prediction towards the majority class.

7.3 Domain-expert analysis

Tab. 4 reports the prevalent patterns of SCADA events that are observed during a PFW or AFW and the corresponding consequent for the association rules having the highest confidence. Within each set (Component/Cause, PFW/AFW), the top 15 rules have been examined, showing the following values of minimum confidence within each subset: Component, PFW - 64.3%; Cause, PFW - 83.6%; Component, AFW - 63.3%; Cause, AFW - 69.4%.
In the case *Component - PFW*, all the examined rules show the conductor as component interested by the PSI. This is quite reasonable, as the majority of protections, switches, as well as breakers are installed for operating them. Besides that, the combinations of the short type of fault and opening of MV line in the antecedent are dominating, suggesting that if the short faults occur more frequently than before, an expected fault will happen in the same monitoring area. In addition, the short circuit fault also appears multiple times in the rules as they would bring relatively high impact on the system. Regarding the case *Cause - PFW*, less meaningful rules are identified. The first type of rule says that given the occurrence of some SCADA events (permanent fault, opening of MV line) some tree may touch a component of the grid, but it is unlikely that the those events are specifically predictive for that kind PSI. The latter rule type is not helpful, since links the occurrence of an electric fault to the absence of any event in the PFW. Even for the case *Component - AFW*, it is not unreasonable that most rules point out to the conductors. Various types of MV line opening may lead to a PSI affecting this type of component. Finally, in the case *Cause - AFW*, it can be observed that when a grounding fault is registered and the automatic reclosing relay (DRA) cannot resolve the

Table 4: Prevalent patterns of SCADA events in the top 15 rules having the highest confidence within each subset (Class label, Window).

Class label	**Window**	**Antecedent**	**Consequent**
Component	PFW	short fault	conductor
		short fault, Opening of MV line (generic)	conductor
		short fault, RG CTO intervention (short circuit)	conductor
Cause	PFW	permanent fault, opening of MV line (various types)	plant fall
		no SCADA events detected	electric fault
Component	AFW	opening of MV line (various types)	conductor
Cause	AFW	opening of MV line (ground), opening of MV line (DRA excluded)	electric fault

problem, this is due to an electric fault. This is coherent with the operation experience, as the DRA is only useful in the case that in a short time the cause of the fault can be self-solve, such as shortly contacting of foreign objects.

8 Conclusions

The work analysed 6 years of data recorded from a medium-voltage distribution network, with the purpose of estimating both the prognostic and diagnostic potential for severe faults, i.e., permanent service interruptions. The proposed approach, consisting of time-window data characterisation, exhaustive rule-mining extraction, and associative classification rules, has been able to assess the potential of the data for fault prognosis and diagnosis. Specifically, the collected SCADA events effectively support the diagnostic task, including the diagnosis of the affected components and fault causes, whereas their prognostic potential is limited since only few and poor predictive correlations are present in the data, and the predictive model based on such rules yielded to very poor results both in recall and precision with unbalanced predictions towards the majority class.

Future works include wider analyses of the rules for different thresholds and changes to the transactional dataset derived from the raw data to enable the extraction of additional correlations.

Acknowledgment

This work has been partially funded by Enel Italia, e-distribuzione, and the SmartData@PoliTO center for Data Science technologies and applications. We are grateful to Paolo Garza and Giuseppe Attanasio for their help in exploiting the L3 associative classifier.

References

1. C. A. Andresen, B. N. Torsaeter, H. Haugdal, and K. Uhlen. Fault detection and prediction in smart grids. In *IEEE 9th International Workshop on Applied Measurements for Power Systems (AMPS)*, pages 1–6, Sep. 2018.
2. Yang Zhang, Tao Huang, and Ettore Francesco Bompard. Big data analytics in smart grids: a review. *Energy Informatics*, 1(1):8, 2018.
3. Chunming Tu, Xi He, Zhikang Shuai, and Fei Jiang. Big data issues in smart grid a review. *Renewable and Sustainable Energy Reviews*, 79:1099 – 1107, 2017.
4. Jian Wang. Early warning method for transmission line galloping based on svm and adaboost bi-level classifiers. *IET Generation, Transmission and Distribution*, 10:3499–3507(8), November 2016.
5. Y. Zhang, Y. Xu, Z. Y. Dong, Z. Xu, and K. P. Wong. Intelligent early warning of power system dynamic insecurity risk: Toward optimal accuracy-earliness tradeoff. *IEEE Transactions on Industrial Informatics*, 13(5):2544–2554, Oct 2017.
6. Q. Cui, K. El-Arroudi, and G. Joos. An effective feature extraction method in pattern recognition based high impedance fault detection. In *19th International Conference on Intelligent System Application to Power Systems (ISAP)*, pages 1–6, Sept 2017.
7. Huaiguang Jiang, Xiaoxiao Dai, Wenzhong Gao, Jun Zhang, Yingchen Zhang, and Eduard Muljadi. Spatial-temporal synchrophasor data characterization and analytics in smart grid fault detection, identification and impact causal analysis. *IEEE Transactions on Smart Grid*, 7:1–1, 09 2016.

8. Enrico De Santis, Lorenzo Livi, Alireza Sadeghian, and Antonello Rizzi. Modeling and recognition of smart grid faults by a combined approach of dissimilarity learning and one-class classification. *Neurocomput.*, 170(C):368–383, December 2015.
9. Y. Cai and M. Chow. Exploratory analysis of massive data for distribution fault diagnosis in smart grids. In *IEEE Power Energy Society General Meeting*, pages 1–6, July 2009.
10. Enrico De Santis, Antonello Rizzi, and Alireza Sadeghian. A learning intelligent system for classification and characterization of localized faults in smart grids. In *2017 IEEE Congress on Evolutionary Computation, CEC 2017, Donostia, San Sebastián, Spain, June 5-8, 2017*, pages 2669–2676, 2017.
11. Rui Zhao, M. R. A. Iqbal, K. P. Bennett, and Qiang Ji. Wind turbine fault prediction using soft label svm. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3192–3197, Dec 2016.
12. S. FajarHazrat, P. Khatri, M. Ahmed Butt, and M. Zama. Grid Monitoring using Solar SCADA Dataset. *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)*, 13(5):39–48, Sep-Oct 2018.
13. N. Bartolini, L. Scappaticci, A. Garinei, Matteo Becchetti, and L. Terzi. Analysing Wind Turbine States and SCADA Data for Fault Diagnosis. *International Journal of Renewable Energy Research*, 7(1):323–329, 2017.
14. Francesco Castellani, Davide Astolfi, and Ludovico Terzi. *Analyzing State Dynamics of Wind Turbines Through SCADA Data Mining*, volume 4, pages 213–223. 01 2016.
15. Li Bin, Zhang Wei-guo, Ning Dong-fang, and Yin Wei. Fault prediction system based on neural network model. pages 496 – 496, 10 2007.
16. J. Liu and G. Geng. Fault prediction for power plant equipment based on support vector regression. In *8th International Symposium on Computational Intelligence and Design (ISCID)*, volume 2, pages 461–464, Dec 2015.
17. R. Han and Q. Zhou. Data-driven solutions for power system fault analysis and novelty detection. In *11th International Conference on Computer Science Education (ICCSE)*, pages 86–91, Aug 2016.
18. D. Liu, B. Wu, C. Gu, Y. Ma, and B. Wang. A multidimensional time-series association rules algorithm based on spark. In *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1946–1952, July 2017.
19. F.-X Qiu, Fengqi Si, and Z.-G Xu. Association rules mining based on principal component analysis and sensor fault detection of power plant. 29:97–102, 02 2009.
20. Xiaoyu Wang, Stephen McArthur, Scott Strachan, John D. Kirkwood, and Bruce Paisley. A data analytic approach to automatic fault diagnosis and prognosis for distribution automation. *IEEE Transactions on Smart Grid*, PP:1–1, 05 2017.
21. Matteo Nisi, Daniela Renga, Daniele Apiletti, Danilo Giordano, Tao Huang, Yang Zhang, Marco Mellia, and Elena Baralis. Transparently mining data from a medium-voltage distribution network: A prognostic-diagnostic analysis. In *Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference, EDBT/ICDT 2019, Lisbon, Portugal, March 26, 2019*, 2019.
22. Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Addison-Wesley, 2005.
23. Bing Liu Wynne Hsu Yiming Ma, Bing Liu, and Yiming Hsu. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, pages 24–25, 1998.
24. E. Baralis, S. Chiusano, and P. Garza. A lazy approach to associative classification. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):156–171, Feb 2008.
25. Robin Genuer, Jean-Michel Poggi, and Christine Tuleau. Random Forests: some methodological insights. Technical report, INRIA, 2008.
26. F. Pedregosa, G. Varoquaux, and A. Gramfort. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.