The Journal of Energy: Engineering & Management
Vol. 11, No. 4, Winter 2022, P. 20-29
Research Article

Association Rule Mining Application to Diagnose Smart Power Distribution System Outage Root Cause

Mohammad Reza Dehbozorgi[1], Mohammad Rastegar[2*]

[1] School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
m.dehbozorgi@shirazu.ac.ir
[2] School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
mrastegar@shirazu.ac.ir

Received: 2/4/2021 **Accepted:** 10/2/2021

Abstract
Smart grids have been introduced to address power distribution system challenges. In conventional power distribution systems, when a power outage happens, the maintenance team tries to find the outage cause and mitigate it. After this, some information is documented in a dataset called the outage dataset. If the team can estimate the outage cause before searching for it, the restoration time will be reduced. In line with smart grid concepts, an association rule-based method is presented in this paper to find the outage cause. To do this, we have first combined outage, load, and weather datasets and extracted features. Then, for every cause, the records are labelled main class or others. The association rules are extracted and evaluated. Through these rules, one can determine whether the outage has happened because of a fault in a certain piece of equipment or not. Doing so alongside using smart devices may lead to reliability enhancement.

Keywords: Power distribution outage, Association rule mining, Distribution systems reliability, Automatic fault management.

* Corresponding author
** This article is a selected article in SGC 2020 Conference, which has been accepted after completion and evaluation.

1. Introduction

With an increase in demands for electricity, the power grid has faced a lot of obstacles that have made it hard to keep reliability at the desired level. The traditional model of power system usually consists of generation level, power transmission, and electricity distribution. As the final part of this scheme, the distribution system plays a pivotal role in the reliability of the system. Also, as a result of load increment, the structure of distribution systemis now more complex than before and needs to be treated with improved management and control techniques. Poor distribution system reliability usually decreases the profit earned by increasing the customers' dissatisfaction. Hence, the smart grid has been proposed as a solution to these problems.

One of the main problems which decrease reliability is power outages that happen because of a fault in a certain piece of equipment. These outages may take time to be found and mitigated because of the complex and radial nature of the feeders. Finding a way to estimate the outage cause before sending a maintenance team can be helpful because the team will only look for the estimated cause when examining the feeder. Also, preventive maintenance (PM) is another way to decrease outage frequency. To improve PM's outcome, one can identify the various factors that affect certain outage cause occurrence and try to decrease their effects. Utilizing the proposed idea in combination with monitoring capabilities in smart grids can decrease repair time and increase reliability.

Recently, many devices have been offered to improve the controllability and reliability of distribution system in form of smart grids. A famous set of devices is usually referred to as system automation. It consists of various control and monitoring instruments such as remote fault locators (RFL), automated switches, and smart relays which are used to control the system remotely. Another scheme is usually used in the form of SCADA systems that gather data from sensors and other measurement instruments to control a certain mechanism or system. Using these devices in the power and distribution systems has led to an enormous volume of valuable raw data. These datasets are usually generated in real or near-real time. There are other data-generating sources such as AMIs that monitor consumed load by the customers. It is interesting to know that as of 2018, 87 million AMI devices have been installed in the US [1]. The size of the mentioned datasets usually complicates the procedures employed to analyze it, and advanced methods should be used to deal with it.

When an outage occurs, a maintenance team is dispatched to find and remove or replace the faulty equipment. When the faulty equipment is taken care of, the service will be restored. These types of outages usually go by the name of permanent outages. After power restoration, information such as the exact time of the outage, the feeder where the fault has occurred, expected energy not supplied (ENS), and the outage cause (the type of the equipment, natural causes, etc.) are documented. There are other types of outages in which the outage duration is below 10 minutes. These outages, called momentary outages, are mitigated either by manual or by automatic reclosing. These records along with respective information are usually registered in certain software programs when documenting the outages. The datasets compiled using these records are called outage datasets in this paper.

To overcome the problems mentioned with the large datasets, the researchers have proposed to use a branch of data science called data mining. It is one of the advanced data analysis techniques that seek to find useful and hidden relationships in a datasets. Some of the frequently used methods include classification, clustering, and association rule mining. Classification models try to build a model called a classifier to diagnose a record's class using a set of features. As an unsupervised technique, clustering aims at finding certain groups in the datasets called clusters where the records are most similar to each other. Usually, the criteria used are records' proximity. Finally, The technique of association rule mining, used here, is employed to find rules that connect a set of items called antecedents to another item entitled the consequent in the dataset.

Recently, several authors have tried to utilize data mining methods to analyse the datasets generated in power systems better. PMUs are special devices that can measure voltage magnitude and phase employing a pre-specified sampling rate. These devices are also capable of sending this information to a control centre. [2] and [3] have proposed to use classification methods to offer models that describe the transient stability mechanism. The authors of [2] have constructed a binary classifier that can tell if the event will jeopardize the system's stability by employing PMU datasets. Also, [3] has suggested using a supporting vector method called CVM can reduce the time spent on constructing the classifier. Using Micro-PMU data to diagnose an event's source is outlined in [4]. This paper shows that doing such will lead to improved situational awareness. Besides, [5] and [6] have proposed methods to obtain approximately the load patterns based on AMI data using classification and the clustering based algorithms respectively. Finally, diagnosing electricity theft employing AMI data is the subject of [7].

Besides using these datasets to analyze the system's behaviour, many have utilized different data-driven methods to improve the reliability of the system. Several tasks can be carried out to enhance reliability including fault locating, outage prediction, and outage cause diagnosis in smart grids.

Fault locating is one of the most vital operations in distribution systems. Sometimes, finding the location of the fault that has caused an outage is a little hard due to an extensive use of radial feeders. To solve this issue, [8]-[9] have used the datasets available to build models that can roughly estimate the fault's location. In [8], the authors have employed Bayesian networks and historical outage information to locate the fault. [10] has recommended applying SVM combined with discrete wavelet transform (DVM) to find the location of the fault in mesh grids. [11] and [9] have also used k-nearest neighbors and rough set theory respectively to locate the element responsible for the outage. In [11], the fault is also classified according to the phases involved (single-phase, three-phase, etc), whereas in [9], the outage cause is identified as well.

Another solution to decrease the effect of outages on the reliability of a system is to predict the time or the number of outages. To do so, several papers have proposed various methods to predict different outage causes. For instance, predicting weather-related outages is addressed in [12] and [13] where the use of random forest and logistic regression (LR) is outlined to find models that can accurately predict weather-caused outages. Other papers including [14] have tried to propose procedures that can estimate the duration of hurricane-caused outages. In addition to the weather's effect, some researchers have taken other causes such as vegetation or birds' activity into account. To predict the outages caused by birds collision, [15] has built a Bayesian network based on weather historical data. Vegetation-related outages prognosis is also discussed in [16] and [17]. [16] solves this problem by proposing to use LR and artificial neural networks (ANN) to find prediction models while [17] does this using time series algorithms.

To increasing reliability, another approach tries to find the outage occurrence or its causes using an outage dataset. For instance, [18] uses SVM to diagnose line outages. The datasets in this paper include the line outage dataset and consumers' meter data. [19] uses ANN and LR to build a model that can tell if the outage is caused by a bird accident or it is a tree-related one. The authors of [20] developed tree classifiers based on decision trees, LR, and naïve Bayesian network where the classes are either equipment related or non-equipment related.

Another data mining method, association rule mining, is a method employed in [21] and [22] to find the important and useful rule that help the operator identify the faulty element faster. [21] proposes five classes--equipment, vegetation, animals, lightning, and public--accidents to describe the outage causes. This paper extracts the rules in which the consequent is one of the classes already mentioned. The antecedents here are a set of features such as temperature, the protective device, etc. [22] solves the same issue, yet it adds a load feature that leads to better and more accurate rules.

As depicted by the previous paragraphs, most of the papers tend to focus on natural causes such as vegetation or animals' activity. In fact, the equipment-related outages have not been investigated comprehensively. Also, how momentary outage occurrence affects the permanent ones and the role of scheduled maintenance have not been considered in similar association rule-mining based papers for identifying the outage cause.

To improve the outage diagnosing, this paper proposes an association rule-mining based method to extract rules that help find the faulty piece of equipment faster. To do this, it processes three different datasets to gain useful features. The first and the most important dataset consists of the permanent outages between Mar. 2015 and Mar. 2019. Several features such as outage month, outage hour, the distance between the permanent outage and the last momentary outage in the same feeder, etc. are calculated for every record. The second dataset indicates the load consumed by the sub-transmission substation that powers the faulty feeder in the form of sub-hourly values recorded by smart metes. We have extracted the average load and normalized average load from this dataset. Finally, the last dataset is weather historical information. Features like temperature, humidity, and wind speed have been obtained. After the pre-processing section, the outage classes are labeled according to the malfunctioning piece of equipment (cable, cable termination, etc.). For every equipment-related outage cause, we have formed a dataset in which the record is labelled as either *main class* or *others* depending on the main outage cause. After balancing, the Apriori algorithm is run, and several rules are obtained. These rules are evaluated using their confidence, rule support, lift, and chi-square to make sure that the most important rules are selected. The results show that by having a handful of features and using one of the extracted rules, one can narrow down the outage causes. If the operators use the estimated cause, they can facilitate the smart grid fault management procedure by informing the maintenance team of the possible outage cause.

The major contributions of this paper are as follows: (i) finding the possible outage cause before sending the

maintenance team which can greatly decrease the repair time. (ii) finding a set of factors that frequently lead to a certain outage cause, which can be used to strengthen the PM's effect on the feeders.

2. Methodology

The use of data mining methods, especially association rules and their benefits, was outlined in the introduction. To apply such a method, first, we will discuss how various datasets are merged to achieve the main dataset that contains useful features. These datasets are outage dataset, sub-transmission substation load, and weather historical data. In the second part, the Apriori algorithm will be introduced and explained. Finally, the proposed method used to extract interesting rules is presented.

2.1. Pre-Processing

As the first step in the pre-processing part, we will define features to describe the outage. The first dataset is called the outage dataset. When an outage occurs, the information regarding the outage is documented. This paper uses the outages that have happened between Mar. 2015 and Mar. 2019. In this dataset, the outages are either scheduled or unscheduled. The available scheduled outages are usually due to preventive equipment maintenance and equipment upgrades. In the current dataset, about 30% of the outages (2032 records) are scheduled outages. The duration of some unscheduled outages is so short that they can be taken care of through manual or automatic reclosing. These outages are called momentary outages in this paper. About 70% of unscheduled outages (3123 records) belong to this category. If the outage is not mitigated by doing the above, the maintenance crew is dispatched to locate the faulty part to localize or repair it. Here, we call these outages as the permanent ones. After power restoration information such as the faulty feeder, the breaking point of the circuit (substation output or a T-off), unsupplied energy, the exact time of the outage, repair duration, the outage cause, etc are logged. Fig.1 depicts the approximate share of the most important causes. Some features are more important than the others in this dataset. For instance, repair duration, ENS, and breaking point do not contain significant info about the outage cause. Hence, they will not be used. To validate this statement a feature selection algorithm was run on this dataset, and it was found out that the most important features are as follows: The month when the outage happens (*OM*); the hour of the outage (*OH*); and the substation's zone. There are seven zones in the present grid that are tags for the sub-transmission substation's location. The demand and the grid configuration are unique to every zone. Hence, they affect some certain outage cause occurrence.

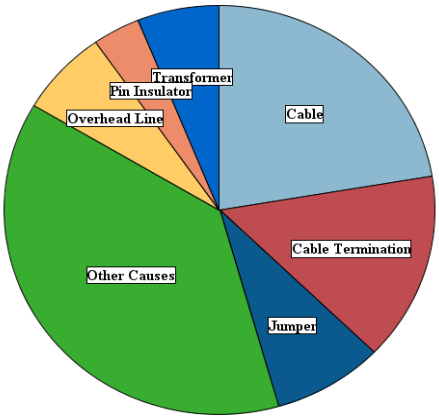

Fig. 1: Share of each of the most important outage causes

Three other features can be generated according to the outage dataset. These features describe the effect of monetary outages and maintenance on permanent outages happening. The first feature is named the number of momentary in the last 30 days (*NM*). It indicates how many short-term outages have happened in the last 30 days in the feeder where the permanent outage has happened. The second feature, days from the last momentary (*DM*) show how many days have passed since the last monetary outage in the feeder. Finally, the last feature, entitled last scheduled maintenance (*LM*), shows the distance between the permanent outage time and last maintenance on the feeder in terms of days.

Another important dataset, especially in the case of equipment faults, is the load dataset. In this paper, the sub-hourly electricity demand for every substation is available for Mar. 2015 to Mar. 2019. Since the load value does not affect outage occurrence instantaneously, it is better to use the average load value of 3 hours for outage occurrence. To do so, the time when an outage happens is obtained. Then, the load level of the sub-transmission substation in the last 3 hours is averaged. Since different feeders are designed for different demand levels, it would be wise to normalize the load value. The average load calculated in the last step is called average load (*AL*). To calculate the normalized load, the peak load of the substation in the respective year is called peak load (*PL*). The normalized load value (*NL*) can be obtained using (1):

$$NL = \frac{AL}{PL} \tag{1}$$

The third dataset is that of historical weather records. In this dataset, retrieved from [23], the values for air temperature (*T*), air humidity (*H*), and wind speed (*WS*) have been documented in the form of 3-hour observations. This dataset covers weather information

from Mar. 2015 to Mar. 2019. To extract the useful features, *OH* (the outage hour) is obtained. The closest weather record to *OH* is used as the weather record for that particular outage. For instance, if the outage has happened at 8 AM and if we have weather reports for 7 AM and 10 AM, the former will be used.

2.2. Apriori Algorithm

In the last section, the pre-processing part was described. In many applications, we need to establish some sort of relationship between the items in hand. For example, in this paper, we wish to find the factors responsible for certain outage causes. In fact, we mean to find the right set of antecedents that leads to the outage cause (consequent). Any dataset can be represented in transactional form, meaning every record (outage) is characterized by a combination of discrete features. The discretisation process will be explained in the next section.

In every transactional database, the first step to find association rules is to extract the most frequent item sets. The item set is usually in the form of *I* in the below notation [24]:

$$I = \{x_1, x_2, ..., x_n\} \quad (2)$$

In the above set, x_n denotes certain items. The numbers are discretisation labels that are defined in section 2.3. The item set shows that there are records that contain these items. In this type of database, the number of records that contain an item set is called the item set's support and is shown by *Sup (I)*.

The simplest method used to mine frequent item sets is to generate a candidate item set. In every stage, item sets with *n* items will be generated. If we take the total number of features used as *m*, the number of candidates in the final stage will be 2^m. After this step, the dataset is scanned to calculate the support of every item set in every stage. The item set with the support value below a pre-set variable called minimum support (*MinSup*) will be rejected as they are not "frequent" [24]. It may take long to generate candidates and much longer to calculate their support. This method is not practical because of the calculational complexity.

The majority of the item sets generated in the last step are infrequent. Thus, an important property called Apriori property indicates that if an item set's support is *s*, its children's support (the item set generated by adding another item to the parent item set) will be either equal or lower than *s* [25]. With this fact in mind, the above algorithm can be modified to reduce greatly the search space and, therefore, the time consumed. To do so, we start at the node where the item set is the null set. In the next stage, we add another item. In this stage, the support of all the item sets will be computed. If it is not frequent, we will not add items to it. In this manner, all the eligible item sets are found.

An association rule is a relationship between a frequent set of items (antecedents) and another item (consequent). This type of rule is in the form of (3). Here, the main focus is to find items (*X*) that lead to a certain outage cause (*Y*).

$$X \rightarrow Y \quad (3)$$

For every rule, we can define some measures that evaluate its precision and interestingness. The first index is the rule's support (*RS*) and is defined by (4). Here, $Sup(X \cup Y)$ denotes the number of records that contain both *X* and *Y*. Also, $|S|$ is the total number of the records. The higher *RS* is, the more general the rule will be. The next criterion, *confidence*, shows the chance of *X* leading to *Y*. *Confidence* is the conditional probability of *Y*'s occurrence given *X* happens. According to this statement, *confidence* is defined by (5). The maximum value for *confidence* is 1. Hence, a higher value of *confidence* means a more precise and more valid rule. Equation (6) defines *lift* as one of the indexes used here. *Lift* is the correlation between *X* and *Y*. Values higher than 1 indicate a positive correlation, whereas values lower than 1 show that they are negatively correlated [25]. If *lift* is higher than 1, the occurrence of *X* leads to the occurrence of *Y*. The higher *lift* value is a sign of a more useful rule and is preferred.

$$RS = \frac{Sup(X \cup Y)}{|S|} \quad (4)$$

$$confidence = \frac{Sup(X \cup Y)}{Sup(X)} \quad (5)$$

$$lift = \frac{Sup(X \cup Y)}{Sup(X)Sup(Y)} \quad (6)$$

The final index used in this research is called the chi-square test. It is a test aimed at determining if two discrete features are related and was first presented by Pearson in 1900 in [26]. It can be used in the rule mining area to find out if *X* and *Y* are indeed related, and, hence, if the rule is important. First, we explain it based on the independence test. It can also be used for association rules evaluation. To use it, a confusion matrix is formed similar to Table 1[27]. Here we want to see if feature #1 (F_1) and feature #2 (F_2) are related. In this table, n_1 and n_2 denote the number of labels for F_1 and F_2 respectively. Also, n_{ij} is the number of records with certain labels for F_1 and F_2. According to the notations introduced in Table 1, the expected frequency for each pair of values is defined by (7). With the expected value for each of the table's cells, we can calculate χ^2 in (8). Next, we need to compute the probability of independence. This

probability (*p*) is calculated through (9). In this equation, *q* is the degree of freedom and is equal to (n_1-1) (n_2-1) [27], and Γ is the sign of the gamma function. If *p* is lower than a variable (*pmax*) set before computation, the two features are dependent. Otherwise, the dependence hypothesis is rejected.

$$e_{ij} = \frac{n_i^1 n_j^2}{|S|} \tag{7}$$

$$\chi^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{8}$$

$$p = \int_{\chi^2}^{\infty} \frac{x^{\frac{q}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{q}{2}}\Gamma(\frac{q}{2})} dx \tag{9}$$

To apply the explained method to association rules we simply change F_1 with *X* and F_2 with *Y*. The difference here is that n_1 and n_2 are both equal to 2. Hence, the outcome for antecedent and the consequent is either true or false. Applying this method clarifies whether the rule's result is indeed related to the real class of the record. By setting a *pmax*, we can find and eliminate the rules which have *p* higher than *pmax*.

2.3. Proposed Method

In the last section, the procedure utilized to gain the final dataset was presented. Apriori algorithm was also outlined and discussed as the base method. Here we propose the technique employed in this paper to extract useful rules.

Table 1: A sample confusion matrix

		F_2				
		a_{21}	a_{21}	…	a_{2n_2}	Row counts
F_1	a_{11}	n_{11}	n_{12}	…	n_{1n_2}	n_1^1
	a_{12}	n_{21}	n_{22}	…	n_{2n_2}	n_2^1
	⋮	…	…	…	…	⋮
	a_{1n_1}	$n_{n_1 1}$	$n_{n_1 2}$	…	$n_{n_1 n_2}$	$n_{n_1}^1$
	Column counts	n_1^2	n_2^2	…	$n_{n_2}^2$	$\|S\|$

The first step to initiate an association rule mining procedure is to discretize the features since item sets should be a combinations of categorical features. To prevent the antecedents in all rules from becoming biased toward a certain feature tag, the equal binning method is employed. It means that the number of records with each tag is equal. Table 2 shows how the continuous features presented before they are discretized and their categorical counterpart.

The main interest here is to find those factors that affect a certain outage cause occurrence. To find these factors it is clear that we have to extract rules in which the consequence is the outage cause. One simple way is to label the records based on the outage cause that we are interested in. Meaning, we can label the important outage cause as *main class*. Hence, other records will be labeled as *others*. This act reduces the number of classes to two. Therefore, the consequent is either *main class* or *others*. Doing so, for every important outage cause creates sub-datasets to the number of important causes.

Clearly, the number of records with the label of *others* is more than *main class* records. If we run the algorithm, the consequence will be *others* for most of the rules. Because of this, balancing is necessary to gain relevant rules. In this paper, the records are balanced in a way that the number of both classes are equal to each other. After the algorithm is run, for every outage cause, the rules in which the consequence is *main class* are separated. Next, these rules are evaluated using the introduced indexes. The index used are confidence, rule support, lift, chi-square test. These values are only defined in the particular sub-dataset. Meaning the rule's support in the original dataset is much lower than in the corresponding subdatasets. Minimum confidence (*mconf*) and minimum rule support (*mrs*) are set to filter out useful rules. Also, as stated earlier *pmax* value is set, and the rules with *p* higher than *pmax* are removed since the relationship between the antecedents and the consequent has not been proved. The rules may have high values for support, confidence, and lift but a low value of *p* because of the balancing process. It should be noted that χ^2 value is calculated over the main imbalanced dataset for every rule.

Table 2: Discretisation procedure used to extract rules

Continues value	Discretised value	Label	Condition
NL	NLD	1	$0 \le NL < 0.373$
		2	$0.373 \le NL < 0.471$
		3	$0.471 \le NL < 0.594$
		4	$0.594 \le NL$
WS (m/s)	WSD	1	$0 \le WS < 2$
		2	$2 \le WS < 3$
		3	$3 \le WS \le 11$
T (C)	TD	1	$-2.2 \le T < 12.6$
		2	$12.6 \le T < 20$
		3	$20 \le T < 29.8$
		4	$29.8 \le T \le 39.8$
H (%)	HD	1	$0 \le H < 17$
		2	$17 \le H < 40$
		3	$40 \le H \le 100$
OM	Season	1	$1 \le OM \le 3$
		2	$4 \le OM \le 6$
		3	$7 \le OM \le 9$
		4	$10 \le OM \le 12$
DM (days)	DMD	1	$0 \le DM < 14$
		2	$14 \le DM < 80$
		3	$80 \le DM < 240$
		4	$240 \le DM$
NM	NMD	1	$NM = 0$
		2	$NM > 0$
LM (days)	LMD	1	$0 \le LM < 41$
		2	$41 <= LM < 98$
		3	$98 \le LM < 232$
		4	$232 \le LM$
OH	Day quarter	1	$0 \le OH < 6$
		2	$6 \le OH < 12$
		3	$12 \le OH < 18$
		4	$18 \le OH < 24$
AL (MW)	ALD	1	$AL < 11.23$
		2	$11.23 \le AL < 15.93$
		3	$15.93 \le AL < 23.07$
		4	$21.06 \le AL \le 38.55$
Substation Zone	Zone	Each substation is assigned to its zone	

3. Results

In this section, we apply the proposed method to find interesting rules that describe the factors affecting certain outage causes occurrence. During the pre-processing part, it was revealed that the majority of outages happen because of malfunction or fault in a certain piece of equipment. There are a lot of pieces of equipment present in the power grid that may be responsible for the outage. Here, some causes are more frequent than others. Hence, it would be better to mine rules about them. These causes include fault in the cable, cable termination, jumper, overhead distribution line, transformers, and pin insulators used in overhead power lines. These rules can be used to either determine the outage cause before searching for it or identify the factors associated with different causes.

As explained in the previous part, the rules in which the consequence is one of the above are of interest. Here we use indexes such as minimum support, minimum confidence, lift, maximum p to evaluate the rules. Given the notations defined in the methodology section we put *mconf*=70%, *mrs*=8%, *pmax*=0.025.

The first cause is the cable. We found out that about 40% of the faults that lead to an outage, happen in the third zone of the grid (Zone=3). The reason behind this is the concentration of underground cables in the area. Because of this fact, it is better to find rules that describe this zone's cable-related outages. Tables 3 and 4 are the rules related to cable outages. In these two tables, we can see that apart from the zone's involvement, other factors such as the season (summer), wind speed, and the distance between the last momentary outage and the present permanent outage are there. It is revealed that high temperature associated with summer increases the chance of a cable-related outage by decreasing the maximum current that the cable can transmit. For other causes, we simply look for those factors in the grid meaning they will not be separated by their zone. In Table 5 the rules to identify features affecting cable-termination fault are presented. The important factors here are season (Autumn) and zones. Besides, we can observe that the feeders in which maintenance has been carried out between 4 to 7 months (LMD=3) before the outage are more likely to be shut down due to a cable-termination fault. Table 6 shows the rule set for the outages that are caused by a jumper fault. By looking at this table, the average load, days from the last momentary outage can be named as the principal features. These rules demonstrate that a momentary outage is likely to lead to a jumper-related permanent

outage. The wind factor can also affect jumper faults by moving them. For the overhead distribution line, the main factors are the load and the temperature according to Table 7.

Table 3: Cable-related outages rules in the grid

Consequent	Antecedents	Support (%)	Confidence (%)	Lift	χ^2	*p*
Cable	Zone = 3 and HD = 2	8.557	92.157	1.772	73.03	Almost zero
Cable	Zone = 3 and WSD = 1	10.57	90.476	1.739	98.09	Almost zero
Cable	Zone = 3 and DMD = 4 and NMD = 1	17.785	89.623	1.723	215.58	Almost zero
Cable	Zone = 3 and Season = 2	8.893	88.679	1.705	77.18	Almost zero
Cable	DMD = 4 and WSD = 1	14.765	77.273	1.486	61.498	Almost zero

High values of load accompanied by high temperature may increase line sag and cause a line to line short circuit. In Table 8, we can see that transformer faults usually happen in springtime in this grid which might have something to do with the winding or transformer insulation structure. At last for pin insulators, as shown in Table 9, these types of faults are highly dependent upon high humidity and the distance between the permanent outage and the momentary one. It is known that humidity facilitates the fault occurrence by providing it with a path for short circuit current. Moreover, some momentary outages can cause partial discharge in the pin insulator and make the next fault a permanent one.

Table 4: Cable-related outages rules in zone 3

Consequent	Antecedents	Support (%)	Confidence (%)	Lift	χ^2	*p*
Cable	Day quarter = 1 and WSD = 1	13.825	93.333	1.647	19.05	Almost zero
Cable	LMD = 2 and DMD = 4	13.825	90.0	1.588	15.739	Almost zero
Cable	TD = 3 and HD = 2 and DMD = 4	10.138	86.364	1.524	8.78	0.003

Table 5: Cable termination-related outages rules

Consequent	Antecedents	Support (%)	Confidence (%)	Lift	χ^2	*p*
Cable termination	Season = 3 and TD = 1 and WSD = 1	9.16	86.111	1.659	35.38	Almost zero
Cable termination	Season = 3 and TD = 1 and NMD = 1	8.906	80.0	1.541	13.79	Almost zero
Cable termination	Zone = 6 and WSD = 1 and NMD = 1	8.397	78.788	1.518	31.98	Almost zero
Cable termination	Zone = 6 and ALD = 2	8.142	75.0	1.445	17.19	Almost zero
Cable termination	TD = 1 and LMD = 3	9.16	72.222	1.391	17.76	Almost zero
Cable termination	LMD = 3 and HD = 3 and NMD = 1	8.142	71.875	1.385	12.5	Almost zero

Table 6: Jumper-related outages rules

Consequent	Antecedents	Support (%)	Confidence (%)	Lift	χ^2	*p*
Jumper	DMD = 1 and HD = 2 and NMD = 2	10.233	81.818	1.557	14.76	Almost zero
Jumper	NLD = 3 and WSD = 3	9.767	80.952	1.54	8.53	0.004
Jumper	ALD = 2 and NMD = 2	9.302	80.0	1.522	6.33	0.012
Jumper	DMD = 1 and ALD = 3	8.372	77.778	1.48	12.26	Almost zero
Jumper	Day quarter = 4 and NLD = 3	10.233	77.273	1.47	11.64	0.001
Jumper	Zone = 7 and DMD = 1	9.767	71.429	1.359	8.81	0.003

Table 7: Overhead distribution line-related outages rules

Consequent	Antecedents	Support (%)	Confidence (%)	Lift	χ^2	*p*
Overhead distribution line	ALD = 4 and TD = 4 and Day quarter = 3	9.551	88.235	1.707	13.51	Almost zero
Overhead distribution line	LMD = 1 and TD = 4	8.989	87.5	1.693	10.56	0.001
Overhead distribution line	NLD = 4 and WSD = 2	8.427	86.667	1.677	5.02	0.025
Overhead distribution line	Zone = 1 and TD = 4	8.989	81.25	1.572	16.65	Almost zero
Overhead distribution line	TD = 4 and Season = 1 and Day quarter = 3	8.989	81.25	1.572	21.17	0.001
Overhead distribution line	DMD = 3 and TD = 4 and NMD = 1	8.427	80.0	1.548	12.09	0.001

Table 8: Transformer-related outages rules

Consequent	Antecedents	Support (%)	Confidence (%)	Lift	χ^2	*p*
Transformer	Season = 1 and Day quarter = 3 and NMD = 1	8.537	85.714	1.654	6.88	0.009
Transformer	Day quarter = 4 and LMD = 1	9.146	80.0	1.544	13.17	Almost zero
Transformer	Season = 1 and Day quarter = 3 and WSD = 3	9.146	80.0	1.544	6.88	0.009
Transformer	Season = 1 and HD = 1 and Day quarter = 3	8.537	78.571	1.516	5.66	0.017

Table 9: Pin insulator-related outages rules

Consequent	Antecedents	Support (%)	Confidence (%)	Lift	χ^2	*p*
Pin insulator	HD = 3 and WSD = 3 and NMD = 2	9.346	90.0	1.965	30.54	Almost zero
Pin insulator	NLD = 2 and DMD = 1	8.411	88.889	1.941	10.76	0.001
Pin insulator	WSD = 1 and DMD = 1	10.28	81.818	1.787	10.59	0.001
Pin insulator	TD = 1 and DMD = 1	9.346	80.0	1.747	10.22	0.001
Pin insulator	TD = 2 and HD = 3 and WSD = 3	9.346	80.0	1.747	16.01	Almost zero
Pin insulator	LMD = 1 and DMD = 1	13.084	78.571	1.716	17.01	Almost zero

4. Conclusion and Future Works

A real outage dataset combined with other real-world datasets was used to improve outage cause diagnosis in the smart grid era. To do this, three datasets were processed to gather important features. Each of the outages was labeled by the faulty piece of equipment to reflect the outage cause. Using the obtained dataset, an Apriori algorithm was utilized to find rules that describe the set of factors that lead to certain causes. To do so, for each cause, we labeled the dataset in a way that the records are either the *main class* or the *others*. Indexes such as rule support, confidence, and chi-square were used to find the most important rules. Distribution system's operators can make use of these rules to instantly estimate the outage cause before looking for the fault cause. In this way, the repair time will be significantly reduced. The rules can also be employed to conduct PM by looking at the factors responsible for certain causes happening.

As future work, other association rule mining algorithms can be used to find more precise rules. Besides, using the dataset which can be collected from advanced meter infrastructure can lead to more features and more precise rules to diagnose outages. Identifying the real reason behind a momentary outage occurrence (the cause) can be a great research ground to prevent the momentary outage from turning into the long permanent ones.

References

[1] "U.S. Energy Information Administration (EIA)", [Online]. Available: https://www.eia.gov/tools/faqs/faq.php?id=108&t=1. [Accessed 13 8 2020].

[2] Guo T. and Milanović J. V., "Online Identification of Power System Dynamic Signature Using PMU Measurements and Data Mining", IEEE Trans, Power Systems, Vol. 31, No. 3, pp. 1760-1768, May. 2016.

[3] Wang B., Fang B., Wang Y., Liu H. and Liu Y., "Power System Transient Stability Assessment Based on Big Data and the Core Vector Machine", IEEE Trans, Smart Grid, Vol. 7, No. 5, pp. 2561 - 2570, Sept. 2016.

[4] Farajollahi M., Shahsavari A., Stewart E. M. and Mohsenian-Rad H., "Locating the Source of Events in Power Distribution Systems Using Micro-PMU Data", IEEE Trans, Power Systems, Vol. 33, No. 6, pp. 6343 - 6354, Nov. 2018.

[5] Chicco G., Napoli R., Piglione F., Postolache P., Scutariu M. and Toader C., "Load pattern-based classification of electricity customers", IEEE Trans, Power Systems, Vol. 19, No. 2, pp. 1232 - 1239, May. 2004.

[6] Chicco G. and Ilie I.-S., "Support Vector Clustering of Electrical Load Pattern Data", IEEE Trans, on Power Systems, Vol. 24, No. 3, pp. 1619 - 1628, Aug. 2009.

[7] Singh S. K., Bose R. and Joshi A., "Energy theft detection in advanced metering infrastructure", in 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, Feb. 2018.

[8] Chien C.-F. Chen S.-L. and Lin Y.-S., "Using Bayesian network for fault location on distribution feeder", IEEE Trans, Power Delivery, Vol. 17, No. 3, pp. 785 - 793, Jul. 2002.

[9] Peng J.-T., Chien C. and Tseng T., "Rough set theory for data mining for fault diagnosis on distribution feeder", IEE Proceedings - Generation, Transmission and Distribution, Vol. 151, No. 6, pp. 689 - 697, Nov. 2004.

[10] Deng X., Yuan R., Xiao Z., Li T. and Wang K. L. L., "Fault location in loop distribution network using SVM technology", International Journal of Electrical Power & Energy Systems, Vol. 65, pp. 254-261, Feb. 2015.

[11] Recioui A., Benseghier B. and Khalfallah H., "Power system fault detection, classification and location using the K-Nearest Neighbors", in 2015 4th International Conference on Electrical Engineering (ICEE), Boumerdes, Dec. 2015.

[12] Wanik D. W., Anagnostou E. N., Hartman B. M., Frediani M. E. and Astitha M., "Storm outage modeling for an electric distribution network in Northeastern USA", Natural Hazards, Vol. 79, p. 1359–1384, Jul. 2015.

[13] Kankanala P., Pahwa A. and Das S., "Regression models for outages due to wind and lightning on overhead distribution feeders", in 2011 IEEE Power and Energy Society General Meeting, Detroit, Jul. 2011.

[14] Nateghi R., Guikema S. D. and Quiring S. M.,

"Forecasting hurricane-induced power outage durations", Natural Hazard, Vol. 74, p. 1795–1811, Jun. 2014.

[15] Sahai S. and Pahwa A., "A Probabilistic Approach for Animal-Caused Outages in Overhead Distribution Systems", in 2006 International Conference on Probabilistic Methods Applied to Power Systems, Stockholm, Jun. 2006.

[16] Radmer D., Kuntz P., Christie R., Venkata S. and Fletcher R., "Predicting vegetation-related failure rates for overhead distribution feeders", IEEE Trans, Power Delivery, Vol. 17, No. 4, pp. 1170 - 1175, Oct. 2002.

[17] Doostan M., Sohrabi R. and Chowdhury B., "A data-driven approach for predicting vegetation-related outages in power distribution systems," International Transactions on Electrical Energy Systems, Vol. 30, No. 1, Aug. 2019.

[18] Hosseini Z. S., Mahoor M. and Khodaei A., "AMI-Enabled Distribution Network Line Outage Identification via Multi-Label SVM", IEEE Trans, Smart Grid, Vol. 9, No. 5, pp. 5470 - 5472, Sept. 2018.

[19] Xu L. and Chow M.-Y., "A classification approach for power distribution systems fault cause identification", IEEE Trans, Power Systems, Vol. 21, No. 1, pp. 53 - 60, Feb. 2006.

[20] Doostan M. and Chowdhury B. H., "Power distribution system equipment failure identification using machine learning algorithms", in 2017 IEEE Power & Energy Society General Meeting, Chicago, Jul. 2017.

[21] Doostan M. and Chowdhury B. H., "Power distribution system fault cause analysis by using association rule mining", Electric Power Systems Research, Vol. 152, pp. 140-147, Nov. 2017.

[22] Bashkari S., Sami A. and Rastegar M., "Outage Cause Detection in Power Distribution Systems based on Data Mining", IEEE Transactions on Industrial Informatics, Vol. 17, No. 1, pp. 640 - 649, Jan. 2021.

[23] "rp5.ru Reliable Prognosis", [Online]. Available: https://rp5.ru/. [Accessed 4 6 2020].

[24] Zaki M. J. and Meira Jr. W., "Chapter 8. Itemset mining", in Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge: Cambridge University Press, 2014, pp. 241-268.

[25] Han J., Kamber M. and Pei J., "Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods", in Data Mining: Concepts and Techniques, Waltham: Morgan Kaufmann, 2012, pp. 243-278.

[26] Pearson K., "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Vol. 50, No. 302, pp. 157-175, 1900.

[27] Zaki M. J. and Meira Jr. W., "Chapter 3. Categorical Attributes", in Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge: Cambridge University Press, 2014, pp. 71-104.