Discussion about the Weather Impact on the Daily Outages in Urban Distribution System

This pre-print refers to the paper presented in 2019 54th International Universities Power Engineering Conference (UPEC), September 3-6, 2019, DOI: 10.1109/UPEC.2019.8893522

Discussion about the Weather Impact on the Daily Outages in Urban Distribution System

Yang Zhang, Andrea Mazza, Ettore Bompard
Department of Energy
Politecnico di Torino
Torino, Italy
{yang.zhang, andrea.mazza, ettore.bompard}@polito.it

Emiliano Roggero, Giuliana Galofaro
Ireti SpA
Gruppo IREN
Torino, Italy
{emiliano.roggero,giuliana.galofaro}@ireti.it

***Abstract*—In this paper, an evaluation approach for analyzing the weather's impact on the number of daily outages in the urban distribution system is explored. By dividing the number of outages into two levels, the task could be carried out as a binary classification problem. In this study, the actual outage data from the distribution system operator is analyzed together with the local weather condition records. First, the tendency of different outage levels to weather conditions is described by the Principal Component Analysis (PCA). Then, the Support Vector Machine (SVM) algorithm is adopted to build the classification model for predicting the outage levels based on the weather condition. An oversampling method is introduced to manage the severe imbalance between the two outage levels. At the end, the performance of the classification model is assessed with the Receiver Operating Characteristic (ROC) curve.**

***Index Terms*—PCA, SVM, distribution system, outages**

XV. Introduction

With the progress of urbanization, the share of population living in the cities is expected to increase up to 68% in 2050 [1]. Correspondingly, the urban power distribution system is facing more challenges for a reliable and secure operation. As the goal of utilities' pursuit, an uninterrupted power supply plays an important role in the customers' satisfaction and cost of operation. Therefore, a prior assessment of the outages becomes a practical problem for the distribution system operators in predictive maintenance and investment evaluation.

In practice, the distribution network is composed of multi-type power feeders and various equipment in a wide area. Consequently, the outages could be induced by various causes, including the defects of equipment, damages from adverse weather conditions, human's error, and so on [2]. In particular in Europe, the underground cables are largely applied in the urban distribution network, which brings difficulties for analyzing the impact from the external factors.

In order to properly address the uncertainty of outages in distribution network, some researches have been carried out to reveal the causes of outages. The authors in [3] have proposed a method to estimate the outage rate of overhead transmission lines under wind storms. In their work, a fragility curve is developed to describe the relationship between the outage rate and the severity of wind storms. However, the wind storms are usually critical to the overhead lines in transmission system and has limited impact on the underground cables in distribution network. The authors in [4] built the weather condition dependent failure rate models based on the high-resolution radar observations of storm characteristics with a Bayesian outage prediction algorithm. As to the outages in distribution network, an ensemble learning approach is introduced in [5] to estimate the weather-caused power outages, especially those caused by the wind and lightning. The relation between duration of unplanned outages in distribution network and environmental factors is analyzed in [6] by learning from historical outage records.

In this paper, our study will focus on the evaluation of daily outages in distribution network considering different weather conditions. The number of outages per day could be divided into two different critical levels: Level I with 0 or only 1 outage in the whole day, while the others are labeled as Level II. Apparently, Level II is more critical to the distribution system operators than Level I. Those days identified as Level II need more attention for predictive maintenance and preparation of a real outage.

The potential impact of weather conditions on these two levels is first tested with the two-dimensional visualization from PCA. As an efficient dimension reduction method, PCA is good at converting the linear dependent dimensions in high dimensional data set into principal components. With more than half of information contained in the first two dimensions, the instances labeled as Level I and Level II could be visualized in a two-dimensional plot. The weather's impact on these two different levels could be roughly analyzed in an intuitive evaluation.

This task is then considered as a binary classification problem with different weather conditions as input and two levels of outage as output. Since most of the days are labeled as Level I in the real-world records, there is a serious imbalance in the data. In order to address this problem, an over-sampling method [7] is adopted to re-balancing the two levels in the data set. This method implements the over-sampling of the minority class by creating "synthetic" examples rather than by sampling with replacement. The synthetic examples are generated based on the original data according to specific rules, which could improve the classifier's performance to some extent. With the re-balanced data set, a classification model is to be built with the SVM algorithm. The ROC curve will be used to verify the validity of the proposed framework for evaluating the weather's severity.

XVI. Data Description

A. Dataset for Analysis

In this paper, the outage data are collected from the local distribution system between the year 2012 and 2017. Among the 6 years' records, the number of days without any outage and the total number of outages in every month are summarized and shown in Figure 5.

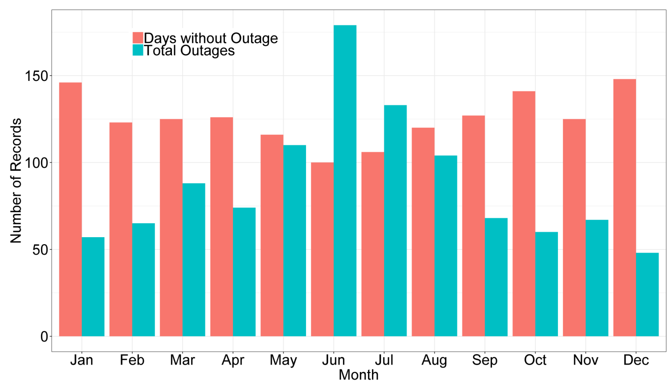

Figure 5 Outage records in different months

In the figure above, the value of each month is the sum of the records in the same month during the 6 years. Accordingly, there are about 180 days of each month. As can be seen from Figure 5, all these months have at least 100 days without any outages, which indicates a reliable and secure power grid. However, a smaller number of days without outages are seen in summer, especially in June and July compared to the winter months. Similarly, the total number of outages in each month reaches the largest value in the summer time.

As for the local weather information, several original data are collected including the daily temperature, relative humidity, precipitation, and so on. The average value of daily maximum temperature in each month is shown in Figure 6.

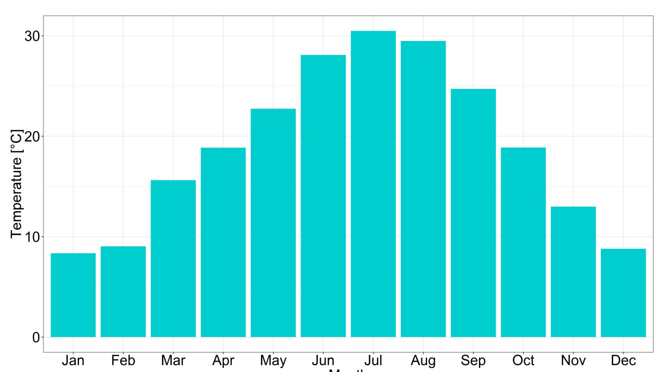

Figure 6 Average value of daily maximum temperature in different months

As is shown in Figure 6, the number of total outages in each month is approximately in cline to the variation of daily maximum temperature in different months. Moreover, the number of days with no outages decreases when the temperature is high in summer.

All these phenomena show a potential relation between the number of outages and the weather conditions. In order to investigate the weather's impact in detail, not only the daily extreme value, but also the average value in a period is taken into consideration for the continuous impact. In this study, apart from the three original daily weather features: maximum temperature, minimum relative humidity and maximum solar radiation, we further calculated their average values in 15 days as three new features for analysis. The daily precipitation and cumulative value of precipitation in 30 days is also constructed as an important feature.

B. Data Visualization

In this study, the severity of outages in distribution system is defined as two levels according to the number of outages per day. Among the outage records in 6 years, the days with more than two independent outages are labeled as "Level II", which indicates a severe situation for special attention, and the rest of days are labeled as "Level I" as shown in Figure 7.

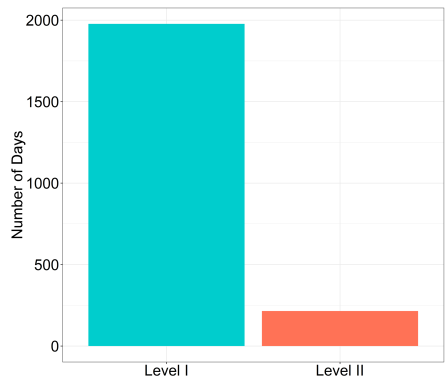

Figure 7 Two levels of daily outages

In order to intuitively demonstrate the weather's impact on the two different outage levels, the PCA method is utilized to reduce the multi-dimensional weather conditions and visualize the two levels on a 2-dimensional plot.

PCA is an orthogonal transformation for dimensionality reduction [8], whose principle is to find a set of optimal base vectors $\mathbf{p}$ to represent the original data $\mathbf{X}$ in the original space as $\mathbf{Y}$ in the new space. In this research, the 8 weather features are regarded as 8 dimensions of the dataset in the original vector space. The covariance matrix of the data in the new space after dimensionality reduction is calculated as follows:

$$\mathrm{Cov}(\mathbf{Y}) = \frac{1}{m}\mathbf{Y}\mathbf{Y}^T = \frac{1}{m}(\mathbf{pX})(\mathbf{pX})^T = \mathbf{p}\left(\frac{1}{m}\mathbf{X}\mathbf{X}^T\right)\mathbf{p}^T \qquad (1)$$

where m is the number of records in our weather dataset.

With an optimal set of base vectors $\mathbf{p}$, the covariance of matrix $\mathbf{Y}$ is supposed to be a diagonal matrix with the covariance of weather features in the new space as 0 and the self-variance as large as possible. It is a classic mathematical problem as orthogonalization of real symmetric matrices. The eigenvalues of the covariance matrix in the original space indicates the percentage of the base vectors (i.e., principal components) in the new space as shown in Figure 8.

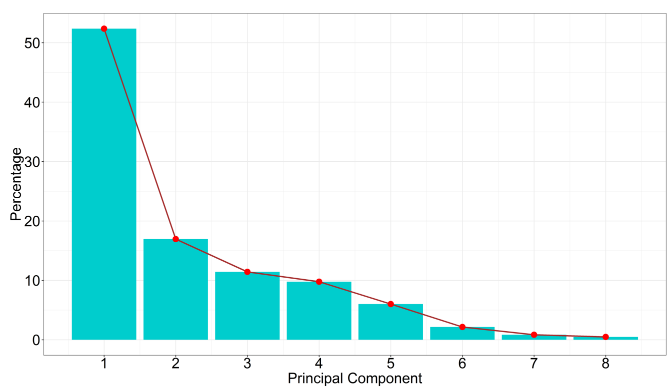

Figure 8 Percentage of principal components

According to the figure above, around 70% information could be represented with the first two largest principal components. Therefore, the daily records during 6 years could be visualized as a 2-dimensional graph in Figure 9 without losing most of the information.

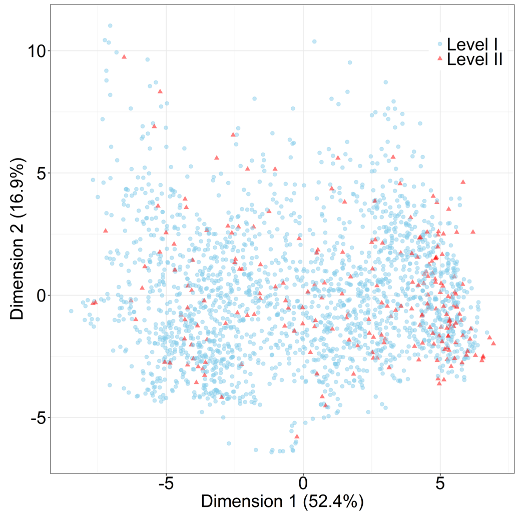

Figure 9 Visualization of two outage levels

As shown in Figure 9, the days defined as outage Level II are labeled in red while the days of Level I are in blue. The spread of red points in general overlaps the scatter of blue points, while the red points are more concentrated on the right part of the graph. Instead, the density of blue points on the right part is not much different to the left part of the graph. This indicates that the outage Level I and Level II could happen in most of the weather conditions and consequently it may be difficult to find a hard boundary purely based on the weather conditions to determine the outage levels. However, the dense red points on the right part of the graph still show that the outage Level II has a higher probability to happen under some certain weather conditions. To reveal the weak relations between the weather conditions and outage levels, a data-driven model is to be built in the following part.

XVII. CLASSIFICATION

As discussed above, the outages in urban distribution network are divided into two levels according to the severity. Therefore, the task becomes a binary classification problem about the outage levels under given weather condition. In this study, the classification model is to be built based on SVM algorithm.

A. SVM Algorithm

SVM is a classic learning algorithm which involves both structural risk minimization principle and statistical learning theory [9]. Given a dataset denoted as $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where *m* is the size of dataset, $\boldsymbol{x}_i$ is a *d*-dimensional vector representing the features of the *i*-th sample and $y_i \in \{-1, +1\}$ as the corresponding class labels. The objective of a binary classification model is to find a mathematical function $h(x_i)$, which satisfies the following equation

$$y_i h(\mathbf{x}_i) = 1 \tag{2}$$

In particular for a linear binary classifier, the above equation can be re-written as below

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) > 0 \tag{3}$$

where the classification model can be expressed as

$$h(\mathbf{x}_i) = sign(\mathbf{w}^T \mathbf{x}_i + b) \tag{4}$$

In an ideal linear binary classification model, the two classes are separated by a hyperplane described with a group of parameters $(\mathbf{w}^T, b)$. The principle of SVM algorithm is to widen the margin between the decision hyperplane and samples for maximizing the generalization capability of the model. Therefore, the mathematical expression of the margin between a data point $\mathbf{p} \in \mathbb{R}^d$ which is the closest one to the hyperplane $(\mathbf{w}^T \mathbf{x}_i + b) = 0$ is used as the objective function in SVM as below:

$$\max_{w,b} \{\min_p (\tfrac{1}{\|\mathbf{w}\|} |\mathbf{w}^T \mathbf{x}_p + b|)\} \tag{5}$$

As can be proven, the scaling of parameters in the hyperplane's expression dose not affect the optimal solution of equation (5). Hence, the basic format of the linear SVM could be expressed as

$$\min_{w,b} (\tfrac{1}{2} \mathbf{w}^T \mathbf{w}) \tag{6}$$

$$\text{s.t.}\quad y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1 \tag{7}$$

In order to avoid the overfitting problem, a slack factor ξ_i is then introduced to allow a small portion of samples classified into the wrong classes via the model, whose expression is as follows:

$$\xi_i = \begin{cases} 0 & if\ y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1 \\ 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) & else \end{cases} \tag{8}$$

Then, the optimization problem in SVM could be formulated as:

$$\begin{cases} \min_{w,b,\xi} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i\right) \\ s.t.\ y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1 - \xi_i \text{ and } \xi_i \geq 0 \\ i = 1, 2, \dots, m \end{cases} \tag{9}$$

where C is the penalty parameter that allows a small portion of misclassification error during the maximization of margin.

According to the optimization theory, the above problem is equivalent to its dual formulation by introducing the Lagrange multiplier α as shown below:

$$\begin{cases} \min_{\alpha} \left(\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j^T - \sum_{j=1}^{m}\alpha_j\right) \\ s.t.\ \sum_{j=1}^{m}\alpha_j y_j = 0 \text{ and } C \geq \alpha_i \geq 0 \\ i = 1, 2, \dots, m \end{cases} \tag{10}$$

The modeling of linear SVM finally becomes an optimization problem for finding a set of optimal parameters $\alpha_i\ (i = 1, \dots, m)$ which satisfy the equation (10). With the obtained values of α, the parameters of hyperplane and could be calculated. For those samples which satisfy the equation (11) are regarded as support vectors.

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1 \tag{11}$$

More generally, for the cases which could not be linearly separated, a kernel function will be introduced to map the original samples to a higher dimensional feature space. One of the typical kernel functions is as in equation (12)

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2} \quad \gamma > 0 \tag{12}$$

B. Over-sampling Technique

As discussed in section II, the two outage levels could be taken as two classes in the binary classification problem. However, there exists a severe imbalance between Level I and Level II. In our study, the synthetic minority over-sampling technique (SMOTE) [7] is utilized to over-sample the minority class (i.e., Level II in our case) for balancing the original samples. The synthetic samples are generated alone the line segments joining any of the k nearest neighbors in the minority class.

The procedures of SMOTE are as follows: For each sample $\mathbf{x}_i$ in the minority class, the *k* nearest neighbors in the same class is located based on the Euclidean distance. Then one of these neighbors $\mathbf{x}_j$ is randomly selected. The synthetic sample $\mathbf{x}_{syn}$ is generated on the line segment of every dimension between the original sample and the selected neighbor according to equation (13).

$$\mathbf{x}_{syn} = \mathbf{x}_i + \delta(\mathbf{x}_i - \mathbf{x}_j) \tag{13}$$

where δ is a random value within [0,1].

C. Results and Performance Evaluation

The SVM-based binary classifier is modeled with the weather conditions and corresponding outage records introduced in Section II. The performance of the model is evaluated upon the test dataset which has not been used for training the model. The initial output is the probabilities of the sample belonging to different classes. With the help of ROC curve as shown in Figure 10, the balance between the True Positive Rate (TPR) and False Positive Rate (FPR) could be achieved with the area under curve as 0.651. Finally, 83% of the samples in Level I and 43% samples in Level II are successfully identified.

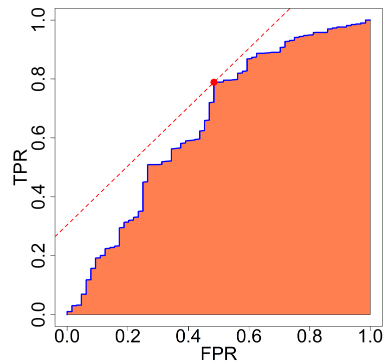

Figure 10 ROC Curve of the SVM Classification Model

XVIII. CLASSIFICATION

In this paper, a data-driven model is discussed to evaluate the impact of weather conditions on the outages in distribution network. Two outage levels are defined according to the number of failure records per day. The weather conditions in this study include the peak values of the day and their average values in 15 days. The results showed a positive but weak relation between the daily outages and weather features considered in this paper. The future work will focus more on the collection of effective features and improve the model for predictive maintenance.

REFERENCES

[1] Department of Economic and Social Affairs, United Nations, “2018 Revision of World Urbanization Prospects”. [online]. Available: https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html

[2] S. Sahai and A. Pahwa, “A probabilistic approach for animal-caused outages in overhead distribution systems”, 9th International Conference on Probabilistic Methods Applied to Power Systems, Stockholm, Sweden, June 11-15, 2016

[3] Y. Liu, S. Lei and Y. Hou, “Overhead transmission line outage rate estimation under wind storms”, IEEJ Transactions on Electrical and Electronic Engineering, vol. 14, pp. 57-66.

[4] M. Yue, T. Toto, M. Jensen, et al, “A Bayesian approach-based outage prediction in electric utility systems using radar measurement data”, IEEE Transactions on Smart Grid, vol. 9, no. 6, 6149-6159, 2018

[5] P. Kankanala, S. Das and A. Pahwa, “Adaboost: an ensemble learning approach for estimating weather-related outages in distribution systems”, IEEE Transactions on Power Systems, vol. 29, no. 1, pp. 359-367, 2014

[6] A. Jaech, B. Zhang, M. Ostendorf, et al, “Real-time prediction of the duration of distribution system outages”, IEEE Transactions on Power Systems, vol. 34, no. 1, pp. 773-781, 2019

[7] N. Chawla, K. Bowyer, L. Hall, et al, “SMOTE: synthetic minority over-sampling technique”, Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002

[8] A. Bosisio, A. Berizzi, D. Le, et al, “Improving DTR assessment by means of PCA applied to wind data”, Electric Power System Research, vol. 172, pp. 193-200, 2019

[9] N. Singh, P. Singh and D. Bhagat, “A rule extraction approach from support vector machines for diagnosing hypertension among diabetes”, Expert Systems with Applications, vol. 130, pp. 188-205, 2019

[10] M. Doostan and B. Chowdhury, “Power distribution system fault cause analysis by using association rule mining”, Electric Power System Research, vol. 152, pp. 140-147, 2017