Electric Power Systems Research 152 (2017) 140–147

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Power distribution system fault cause analysis by using association rule mining

Milad Doostan*, Badrul H. Chowdhury

Electrical and Computer Engineering Department, University of North Carolina at Charlotte, Charlotte, USA

ARTICLE INFO

Article history:
Received 7 March 2017
Received in revised form 5 July 2017
Accepted 6 July 2017

Keywords:
Apriori algorithm
Association rule mining
Fault cause analysis
Machine learning
Pattern recognition
RandomForestSRC
ROSE
Synthetic data
SMOTE

ABSTRACT

In recent years, with the increasing requirements on power distribution utilities to ensure system reliability and to improve customers and regulators satisfaction, utilities seek to find practical solutions that enable them to restrict specific faults or to better manage their responses to unavoidable power outages. For achieving either, it is crucial to acquire a profound understanding of different faults by exploring their underlying causes and identifying key variables related to those causes. Currently, statistical models as well as advanced data analytics techniques are common tools to gain such understanding. Although basic statistical analysis provides a general knowledge of the primary causes of faults; nevertheless, it falls short of describing nuanced conditions that lead to a fault. On the other hand, applying sophisticated algorithms can produce deeper insight into the main causes; however, it would be computationally burdensome and might require a tremendous amount of running time. In order to overcome these problems, this paper proposes a novel approach for fault cause analysis by using association rule mining. The primary goals are to characterize faults according to their underlying causes and to identify important variables that strongly impact fault frequency. This paper proposes a step-by-step procedure, which deals with data preparation, practical issues associated with fault data sets, and implementation of association rule mining. The procedure is followed by a comprehensive case study to demonstrate how the proposed approach can be used to mine for causal structures and identify frequent patterns for vegetation, animal, equipment failure, public accident, and lightning-related faults.

1. Introduction

1.1. Motivation

Faults in power distribution systems can seriously endanger the system operation in different ways. As a matter of fact, distribution faults negatively influence the system reliability since they are responsible for a considerable number of major interruptions that customers experience [1]. Furthermore, faults exert damaging impacts on system safety and security and result in heavy costs for distribution utilities [2]. Therefore, utilities either seek to find practical solutions aimed at preventing specific faults or attempt to take effective measures to properly and quickly restore the system after faults occur. For attaining either, it is essential to acquire a deeper understanding of the primary causes of faults and to identify significant variables related to those causes.

Power distribution utilities are usually interested in preventing avoidable faults. One main course of action to fulfill this objective is to introduce necessary modifications to system fault management based on the knowledge acquired from faults that have occurred in the past. For instance, if it turns out that a distribution system had experienced most of its vegetation-related faults during the months of June and July, then the utility will recognize the need to carry out essential preventive maintenance for those two months. In addition to modifying fault management practices, such knowledge is highly beneficial for improving the design of existing distribution systems to reduce the number of future faults [3]. In order to gain this knowledge, it is required to carry out an in-depth root cause analysis for different faults.

Nevertheless, most often, distribution faults have proved to be unavoidable. For example, even with the implementation of different preventive measures such as installing animal guards or designing large clearances between phase and ground wires, substantial numbers of animal-related faults occur in the system. In these cases, distribution utilities attempt to take an appropriate response to the fault either by predicting it or identifying the causes immediately after the fault. The task of predicting or identify-

* Corresponding author.
E-mail addresses: mdoostan@uncc.edu (M. Doostan), b.chowdhury@uncc.edu (B.H. Chowdhury).

http://dx.doi.org/10.1016/j.epsr.2017.07.005
0378-7796/

ing faults has been considered extremely challenging due to the random nature of faults and the numerous contributing factors. However, over the recent years, with the explosion in data gathering within the smart grid framework, applications of advanced data analytics techniques combined with the traditional rigorous mathematical modeling have facilitated the tasks of fault prediction and identification [4]. In fact, these applications have been identified as viable technologies with the capability to provide necessary actionable information support for utilities to predict or identify different faults with satisfactory accuracy [5–7]. For the purpose of developing these models, however, it is crucial to identify and utilize the factors that are related to each fault cause. For example, the authors in [6] consider six features of: weather condition, season, time of day, faulty phases, protective device activated, and the location where the fault happened as the most influential factors for vegetation and animal-related faults and use them as inputs to build a power distribution fault cause identifier.

Consequently, characterizing faults according to their underlying causes and identifying significant variables that strongly impact the fault frequency are extremely valuable as they allow utilities to find solutions to restrict specific causes and to give an appropriate response to unavoidable faults.

1.2. Literature review

By this time, several studies have been conducted to analyze the characteristics of various faults, which have been caused primarily by animals or vegetation-related issues. These studies mainly take advantage of statistical techniques and data analytics algorithms.

For instance, the authors in [3] propose a statistical data mining approach to perform a root cause analysis of animal-related faults. In their work, the impact of six factors of weather condition, season, the day of the week, time of day, faulty phases, and protection device activated is considered to be significant on the frequency of such faults. In another attempt, by utilizing four statistical measures of actual, normalized, relative, and likelihood, the authors in [8] investigate tree-related faults with respect to six factors: weather condition, season, time of day, the number of faulty phases, location and the clearing device. Furthermore, the same authors in [9] use logistic regression to explore the influence of the aforementioned factors on tree faults by evaluating significance levels resulting from regression. In their study, the weather condition is observed to be the most influential factor. In [10], the authors review two statistical methods, namely hypothesis test, and stepwise regression and introduce two new methods to select proper features for identifying root causes of different faults. They employ these methods to examine six factors explained in [6].

1.3. Contributions

Although basic statistical analysis provides a general understanding of the primary causes of faults, it falls short of describing nuanced conditions that lead to a fault. In fact, gaining a deeper understanding of the contributing factors for each fault type by using conventional statistical methods could become extremely time consuming as it requires performing various analyses. On the other hand, applying sophisticated methods such as stepwise and logistic regression or artificial neural networks can produce deeper insight into the underlying causes; however, it would be computationally burdensome and might require a tremendous amount of running time [10]. The reason for the extra computation time requirement is that a substantial number of inputs, which are based on the many attributes that are involved in the study, have to be generated and fed into these methods.

In order to overcome the aforementioned problems, this paper proposes an approach for fault cause analysis based on association rule mining. Association rules, introduced by [11], belong to a category of uncomplicated but remarkably powerful regularities in binary data. They initially were employed to mine large collections of basket data type transactions for association rules between sets of items. However, due to their proven capabilities for finding interesting associations, or correlations among data items, they have received increasing attention in different fields for practical applications. The main advantage of association rule mining over the aforementioned methods is that it can easily analyze the co-occurrence of different attributes to find any association or correlation. Also, since association rule algorithms were originally developed to be applied to extremely large transaction data sets, they are very effective with regard to the computation time requirement. In fact, as will be demonstrated in the following sections of this paper, the association rule mining is capable of extracting comprehensive patterns from fault data sets within a short amount of time. To implement it, however, in-depth data preparation is required. Moreover, there are several practical issues associated with fault data sets that ought to be addressed. Therefore, this paper provides a step-by-step procedure that fully deals with necessary data preparation, practical issues associated with fault data sets, and implementation of association rule mining. Furthermore, this procedure is applied to investigate a real-world fault data set. As a result of the case study, causal structures and frequent patterns for vegetation, animal, equipment failure, public accident, and lightning-related faults, which have drawn less attention previously in the literature are explored.

1.4. Paper organization

This paper is organized as follows. Association rule mining is discussed in Section 2. In Section 3, the problem of data insufficiency is explained, and a practical solution to deal with it is provided. The proposed methodology is elaborated in Section 4. Section 5 provides a case study to illustrate the implementation of the proposed methodology. Finally, Section 6 draws conclusions on the effectiveness of association rule mining in fault cause analysis.

2. Association rule mining

As mentioned earlier, association rules were initially introduced to mine large collections of basket data type transactions to investigate how items or objects are related to each other. However, nowadays, because of their effectiveness, they are widely employed in different domains to mine for causal structures and to identify frequent patterns in various data sets.

An association rule is a causality, where a rule is defined as an implication of the form $X \Rightarrow Y$, with two conditions of $X, Y \subseteq I$ and $X \cap Y = \emptyset$, where I is a set of n binary attributes called items [11]. The itemsets X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule, respectively.

Given the set of transactions, numerous rules can be generated. However, the rules that are of actual interest to data analysts, which provide useful information, have to fulfill certain constraints. The major constraints related to the “support” and “confidence” of a rule [11]. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X. Support and confidence can be mathematically expressed as (1) and (2), respectively.

$$Support(X \Rightarrow Y) = \frac{\sigma(X, Y)}{N} \tag{1}$$

$$Confidence(X \Rightarrow Y) = \frac{\sigma(X, Y)}{\sigma(X)} \tag{2}$$