Electric Power Systems Research 140 (2016) 413–423

Contents lists available at ScienceDirect

Electric Power Systems Research

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

Unsupervised non-technical losses identification through optimum-path forest

Leandro Aparecido Passos Júnior[a], Caio César Oba Ramos[c,*], Douglas Rodrigues[a], Danillo Roberto Pereira[b], André Nunes de Souza[c], Kelton Augusto Pontara da Costa[b], João Paulo Papa[b]

[a] Department of Computing, Federal University of São Carlos, São Carlos, Brazil
[b] Department of Computing, São Paulo State University, Bauru, Brazil
[c] Department of Electrical Engineering, São Paulo State University, Bauru, Brazil

ARTICLE INFO

Article history:
Received 7 January 2016
Received in revised form 24 May 2016
Accepted 31 May 2016
Available online 28 June 2016

Keywords:
Non-technical losses
Optimum-path forest
Clustering
Anomaly detection

ABSTRACT

Non-technical losses (NTL) identification has been paramount in the last years. However, it is not straightforward to obtain labelled datasets to perform a supervised NTL recognition task. In this paper, the optimum-path forest (OPF) clustering algorithm has been employed to identify irregular and regular profiles of commercial and industrial consumers obtained from a Brazilian electrical power company. Additionally, a model for the problem of NTL recognition as an anomaly detection task has been proposed when there are little or no information about irregular consumers. For such purpose, two new approaches based on the OPF framework have been introduced and compared against the well-known *k*-means, Gaussian mixture model, Birch, affinity propagation and one-class support vector machines. The experimental results have shown the robustness of OPF for both unsupervised NTL recognition and anomaly detection problems. In short, the main contributions of this paper are fourfold: (i) to employ unsupervised OPF for non-technical losses detection, (ii) to model the problem of NTL as being an anomaly detection task, (iii) to employ unsupervised OPF to estimate the parameters of the Gaussian distributions, and (iv) to present an anomaly detection approach based on unsupervised optimum-path forest.

1. Introduction

Non-technical losses (NTL) identification has been paramount in the last years, mainly in development countries where the energy taxes are based on the amount of thefts in the distribution system. The main concept related to NTL refers to the amount of energy that has been used but not billed, and may also reflect on the quality of distributed energy. Aiming to avoid thefts in power distribution systems, or even cyber-attacks [1], one can refer to a considerable collection of works that deal with NTL identification using machine learning techniques.

In regard to supervised-based NTL identification, Nagi et al. [2], for instance, proposed a hybrid approach composed of a support vector machines (SVMs) classifier and fuzzy models that outperformed their previous work, which employed only SVMs for this task [3]. Later, the same group of authors presented an interesting work comparing several supervised learning techniques for NTL detection [4]. Ramos et al. [5] introduced the supervised optimum-path forest (OPF) classifier in the context of theft detection in Brazilian consumers, obtaining results comparable to the ones achieved by SVMs, but much faster considering the training and testing phases. Further, Ramos et al. [6,7] presented two feature selection techniques for NTL characterization, i.e., the authors were concerned into finding the set of features that best discriminate legal and illegal profiles.

Since NTL-oriented datasets are usually imbalanced, Martino et al. [8] presented an approach to deal with such shortcoming, given most part of datasets usually contain much less positive (theft) samples than regular consumers. León et al. [9] presented an expert system for supervised NTL identification in Spain, and the aforementioned group of authors have tackled the same problem through regression analysis [10]. In addition, the MIDAS system was proposed by Monedero et al. [11] aiming to combat thefts in electrical power distribution systems based on neural networks and statistical analysis. Recently, Guerrero et al. [12] used a knowledge-based system that employs text mining,

* Corresponding author at: Av. Eng. Luiz Edmundo C. Coube, 14-01, 17033-360 Bauru, SP, Brazil. Tel.: +55 14 31036115.
E-mail addresses: leandropassosjr@gmail.com (L.A. Passos Júnior), papa@fc.unesp.br (J.P. Papa).

http://dx.doi.org/10.1016/j.epsr.2016.05.036
0378-7796/

neural networks, and statistical techniques to detect non-technical losses.

However, most part of works address the problem of NTL identification using supervised techniques. Since to obtain labelled datasets in this context is usually a hard task, it is often necessary to evaluate unsupervised techniques for NTL identification. Donadel et al. [13], for instance, presented a methodology to estimate the energy consumption based on clustering and sampling techniques, and Tasić and Stojanović [14] employed a fuzzy-based clustering algorithm for energy losses identification. In India, Babu et al. [15] aimed at identifying NTL and suspect profiles of irregular energy consumption by determining similar groups through fuzzy c-means clustering.

Recently, Rocha et al. [16] proposed an unsupervised version of the OPF classifier, in which the pattern recognition task is modelled as a graph partition problem where some key samples (prototypes) compete among themselves in order to partition the graph into clusters. Such approach has been used in several applications, but it has been noticed only one work that employed OPF for unsupervised NTL identification so far [17]. Therefore, this work aims at evaluating OPF clustering for non-technical losses identification in power distribution systems in two private datasets provided by a Brazilian electrical power company, being one dataset composed of commercial profiles and the another one composed of industrial consumers.

Another main contribution of this paper is to model the problem of non-technical losses identification as an anomaly detection task, in which the classifier is trained with the regular consumers only. Therefore, when a new sample comes up to be classified, it is identified if this sample belongs to the "normal" pattern learned by the classifier. Otherwise, such sample is then classified as being an anomaly, i.e., a profile from a thief user. One of the most commonly used approaches to detect anomalies is the so-called Multivariate Gaussian Distribution, in which each cluster is modelled as a Gaussian distribution with its parameters estimated by some machine learning technique. After that, when a new test sample appears to be classified, it is verified its closest Gaussian distribution, and if this distance is greater than a threshold, such sample is then classified as an anomaly. In this paper, the unsupervised OPF has been introduced to estimate the Gaussian parameters, showing it can be more robust than some techniques that are often used for such purpose. Additionally, OPF is compared against the well-known k-means, Birch, affinity propagation (AP), and Gaussian mixture model (GMM) considering both approaches, i.e., unsupervised NTL recognition and anomaly detection (in regard to this last task, one-class SVMs [18] has also evaluated either; and AP and Birch have been applied to unsupervised NTL identification only). Finally, an anomaly detection approach purely based on the optimum-path forest has been proposed as well, which was evaluated in the context of NTL identification.

In short, the main contributions of this paper are fourfold: (i) to employ unsupervised OPF for non-technical losses detection, (ii) to model the problem of NTL as being an anomaly detection task, (iii) to employ unsupervised OPF to estimate the parameters of the Gaussian distributions, and (iv) to present an anomaly detection approach based on unsupervised optimum-path forest. The remainder of this paper is organized as follows: Sections 2 and 3 present the OPF clustering theory and the experimental methodology, respectively. Section 4 discusses the experiments, and Section 5 states conclusions and future works.

2. Optimum-path forest clustering

Let $\mathcal{Z}$ be a dataset such that for every sample $s \in \mathcal{Z}$ there exists a feature vector $\vec{v}(s)$. Let $d(s, t)$ be the distance between s and t in the feature space. For instance, $d(s, t) = \|\vec{v}(t) - \vec{v}(s)\|$ – the Euclidean distance between $\vec{v}(t)$ and $\vec{v}(s)$. A graph $(\mathcal{Z}, \mathcal{A}_k)$ can be defined such that the arcs $(s, t) \in \mathcal{A}$ connect k-nearest neighbours in the feature space. The arcs are weighted by $d(s, t)$ and the nodes $s \in \mathcal{Z}$ are weighted by a probability density value $\rho(s)$:

$$\rho(s) = \frac{1}{\sqrt{2\pi\sigma^2}|\mathcal{A}_k(s)|} \sum_{\forall t \in \mathcal{A}_k(s)} \exp\left(\frac{-d^2(s,t)}{2\sigma^2}\right), \quad (1)$$

where $|\mathcal{A}_k(s)| = k$, $\sigma = d_f/3$, and d_f is the maximum arc weight in $(\mathcal{Z}, \mathcal{A}_k)$. This parameter choice considers all adjacent nodes for density computation, since a Gaussian function covers most samples within $d(s, t) \in [0, 3\sigma]$. Moreover, since $\mathcal{A}_k$ is asymmetric, symmetric arcs must be added to it on the plateaus of the probability density function (pdf) in order to guarantee a single root per maximum.

The traditional method to estimate a pdf is by Parzen-window. Eq. (1) can provide the Parzen-window estimation based on an isotropic Gaussian kernel when the arcs are defined by $(s, t) \in \mathcal{A}_k$ if $d(s, t) \leq d_f$. However, this choice presents problems with the differences in scale and sample concentration. Solutions for this problem lead to adaptive choices of d_f depending on the region of the feature space. By taking into account the k-nearest neighbours, the method handles different concentrations and reduces the scale problem to the one of finding the best value of k, say k^* within $[k_{\min}, k_{\max}]$, for $1 \leq k_{\min} < k_{\max} \leq |\mathcal{Z}|$.

The solution proposed by Rocha et al. [16] to find k^* considers the minimum graph cut among all clustering results for $k \in [1, k_{\max}]$ ($k_{\min} = 1$), according to the normalized measure $GC(\mathcal{A}_k, L, d)$ suggested by Shi and Malik [19]:

$$GC(\mathcal{A}_k, L, d) = \sum_{i=1}^{c} \frac{W_i'}{W_i + W_i'}, \quad (2)$$

$$W_i = \sum_{\forall (s,t) \in \mathcal{A}_k | L(s) = L(t) = i} \frac{1}{d(s,t)}, \quad (3)$$

$$W_i' = \sum_{\forall (s,t) \in \mathcal{A}_k | L(s) = i, L(t) \neq i} \frac{1}{d(s,t)}, \quad (4)$$

where $L(t)$ is the label of sample t, W_i' uses all arc weights between cluster i and other clusters, and W_i uses all arc weights within cluster $i = 1, 2, \ldots, c$.

The method defines a path π_t as a sequence of adjacent samples starting from a root $R(t)$ and ending at a sample t, being $\pi_t = \langle t \rangle$ a trivial path and $\pi_s \cdot \langle s, t \rangle$ the concatenation of π_s and arc (s, t). It assigns to each path π_t a value $f(\pi_t)$ given by a connectivity function f. A path π_t is considered optimum if $f(\pi_t) \geq f(\tau_t)$ for any other path τ_t.

Among all possible paths π_t from the maxima of the pdf, the method assigns to t a path whose minimum density value along it is maximum. That is, the method finds $V(t) = \max_{\forall \pi_t \in (\mathcal{Z}, \mathcal{A}_k)} \{f(\pi_t)\}$ for $f(\pi_t)$ defined by:

$$f(\langle t \rangle) = \begin{cases} \rho(t) & \text{if } t \in \mathcal{R} \\ \rho(t) - \delta & \text{otherwise,} \end{cases}$$
$$f(\langle \pi_s \cdot \langle s, t \rangle \rangle) = \min\{f(\pi_s), \rho(t)\}, \quad (5)$$

for $\delta = \min_{\forall (s,t) \in \mathcal{A}_k | \rho(t) \neq \rho(s)} |\rho(t) - \rho(s)|$ and $\mathcal{R}$ being a root set, discovered on-the-fly, with one element per each maximum of the pdf. It should be noted that higher values of δ reduce the number of maxima. In this work, we used $\delta = 1.0$ and $\rho(t) \in [1, 1000]$. The OPF algorithm maximizes the connectivity map $V(t)$ by computing an optimum-path forest – a predecessor map P with no cycles that assigns to each sample $t \notin \mathcal{R}$ its predecessor $P(t)$ in the optimum path from $\mathcal{R}$ or a marker *nil* when $t \in \mathcal{R}$. Algorithm 1 implements this procedure.