On the Study of Commercial Losses in Brazil: A Binary Black Hole Algorithm for Theft Characterization

Caio C. O. Ramos, Douglas Rodrigues, André N. de Souza, *Member, IEEE*, and João P. Papa, *Member, IEEE*

***Abstract*—According to The Brazilian Electricity Regulatory Agency, Brazil reached a loss of approximately U.S.$ 4 billion in commercial losses during 2011, which correspond to more than 27 000 GWh. The strengthening of the smart grid has brought a considerable amount of research that can be noticed, mainly with respect to the application of several artificial intelligence techniques in order to automatically detect commercial losses, but the problem of selecting the most representative features has not been widely discussed. In this paper, we make a parallel among the problem of commercial losses in Brazil and the task of irregular consumers characterization by means of a recent meta-heuristic optimization technique called Black Hole Algorithm. The experimental setup is conducted over two private datasets (commercial and industrial) provided by a Brazilian electric utility, and it shows the importance of selecting the most relevant features in the context of theft characterization.**

***Index Terms*—Commercial losses, black hole algorithm, optimum-path forest.**

I. Introduction

THE ENERGY losses are defined as the difference between the energy generated or purchased and the energy billed, being classified in two different types: technical and commercial losses. The former are related to the physical characteristics of the energy system, i.e., the technical losses are defined as the energy lost in the transportation, transformation and in the measuring equipments, being its costs predicted by the electric utilities [1]. The commercial losses, also called non-technical losses (NTL), are associated with the energy delivered to the consumer that is not billed, being more difficult to be detected and quantified.

Manuscript received October 17, 2015; revised April 4, 2016; accepted April 18, 2016. Date of publication April 29, 2016; date of current version February 16, 2018. This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, in part by the Fundação de Amparo à Pesquisa do Estado de São Paulo under Grant 2014/16250-0, and in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico under Grant 306166/2014-3. *(Caio C. O. Ramos and Douglas Rodrigues contributed equally to this work.)* Paper no. TSG-01348-2015. *(Corresponding author: Caio C. O. Ramos.)*

C. C. O. Ramos and A. N. de Souza are with the Department of Electrical Engineering, São Paulo State University, Bauru 17033-360, Brazil (e-mail: caioramos@gmail.com).

D. Rodrigues is with the Computer Science Department, Federal University of São Carlos, São Carlos 13565-905, Brazil.

J. P. Papa is with the Department of Computing, São Paulo State University, Bauru 17033-360, Brazil.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2016.2560801

The problem of commercial losses is not faced only in emerging countries, but worldwide, even in smaller proportions. Some experts tend to correlate commercial losses with the development of a given country, which may also include aspects of education, income distribution and violence, among others. The percentage for commercial losses rates vary between 0.5% and 25% in Brazil, 20% and 40% in India, 0.2% and 1% in United Kingdom and around 3.5% in Philippines [2]–[5].

Brazil, for instance, is the largest economy in Latin America and the seventh in the world, being also a member of the BRICS group (Brazil, Russia, India, China and South Africa) that stands for the most prominent developing countries, in other words, they refer to the large emerging markets. It is clear that, as the Brazilian economy grows, the consumption of electricity also increases, bringing together very high losses rates (around 17.5%) [6].

Lately, the problem of detecting commercial losses in power systems is a topic that has been extensively researched [2], [7], [8]. Theft and tampering of power meters in order to adulterate the measurement of power consumption are the main causes that lead to commercial losses in electric power companies [9]. Additionally, performing periodic inspections to minimize such frauds can become very costly in some cases, considering that it is a difficult task to calculate or even measure the amount of losses, and in most cases is almost impossible to know where they occur [10].

Aiming to reduce fraud and electricity theft, several electric power companies have been interested to characterize a profile of irregular consumers, which are mainly related to the illegal electrical installations. The minimization of losses can assure investments in quality programs of energy, as well as it can allow a reduction of the final energy price to the consumer. Nowadays, some advances in this research field can be noted with the use of several artificial intelligence techniques in order to automatically detect commercial losses, which is a real application in smart grids. It is also important to note that the commercial losses are a global issue and the solution to this problem is not trivial [2], [10].

Despite the extensive use of machine learning techniques for the detection of commercial losses in power systems, the problem of selecting the most representative features has not been widely discussed in the context of commercial losses [11]–[15]. In regard to the cutting edge research in this field, Nizar *et al.* [16] proposed a work to select a

subset of samples in order to improve the identification of irregular consumers, and Ramos *et al.* [17] proposed a new hybrid feature selection algorithm based on Harmony Search and Optimum-Path Forest. Further, Ramos *et al.* [18] presented a new methodology based on evolutionary algorithms to the same purpose. Therefore, to point the subset of the most discriminative features to design more effective systems for commercial losses detection is as important as the detection of such losses.

In the last years, there has been an increasing number of researches that concern the problem of feature selection as an optimization task, which can be addressed by means of meta-heuristic and swarm-based techniques. There are a plenty of them, being the most known the Particle Swarm Optimization (PSO) [19], Differential Evolution (DE) [20], Genetic Algorithm (GA) [21] and Harmony Search (HS) [22], among others. The "No Free Lunch Theorem" [23], which states there is not a single optimization approach that outperforms another one for all optimization problems, may contribute with the development of such new algorithms every time.

Recently, an interesting approach presented by Hatamlou [24] for data optimization called Black Hole Algorithm (BHA), that is based on the formation of the well-known black holes in the universe and their attraction power. This approach has demonstrated interesting results in the context of continuous-valued optimization problems. However, to the best of our knowledge, only we have been applied the new technique based on BHA to the context of feature selection, although a binary-constrained optimization version of BHA have been presented by Nemati *et al.* [25] as well.

This paper brings the problem of commercial losses in Brazil and to highlight the use of intelligent computational tools by electric power companies aiming the revenue recovery. Therefore, the main contributions of this paper are two fold: (i) to shed light over the problem of commercial losses in Brazil focusing on the past years, as well as (ii) to present a novel binary optimization algorithm based on the BHA for irregular consumers characterization. The remainder of this paper is organized as follows. Section II presents the context of commercial losses in Brazil, and Section III states the theoretical aspects of BHA. Section IV presents a case study with respect to theft characterization, and finally the conclusions are presented in Section VI.

II. Commercial Losses in Brazil

In Brazil, according to ANEEL (The Brazilian Electricity Regulatory Agency) [3], the losses with energy theft (irregular consumption) have reached the level of R$ 8.1 billion (approximately U.S.$ 4 billion) per year, considering 61 of the 63 utilities who passed the second tariff review cycle in the period from 2007 to 2010. In terms of energy, this amount corresponds to more than 27,000 Gigawatt-hours (GWh), approximately 8% of the consumption in the Brazilian energy captive market, which is formed by consumers that can only buy energy from a distribution utility that operates in the network they are connected.

TABLE I
Amount of Commercial Losses in Each Region of Brazil (Source: ANEEL-2011)

Placing	Region	Commercial Losses (%)
1	North	20%
2	Southeast	10%
3	Northeast	9%
4	Midwest	5%
5	South	3%

TABLE II
Commercial Losses Rates of Electric Utilities in Brazil (Source: ANEEL-2011)

Placing	Utilities	Commercial Losses (%)
1	CELPA	24.4%
2	LIGHT	24.2%
3	CERON	22.0%
4	CEMAR	17.8%
5	AMPLA	17.1%
6	CEAL	17.0%
7	AMAZONAS ENERGIA	16.8%
8	ELETROACRE	15.9%
9	CEPISA	15.8%
10	ENERGISA PARAÍBA	11.2%
11	ELETROPAULO	10.8%
12	CEEE	10.5%
13	BANDEIRANTE	10.1%
14	ESCELSA	10.0%
15	BOA VISTA	10.0%

In order to clarify the problem of energy theft in Brazil, Table I [3] shows the amount of commercial losses for each Brazilian region. The largest amount of losses can be observed in North region, where the implementation of procedures for handling technical and commercial losses are not easy to be performed, mainly due to the difficult access and the large territory in which the electric utility operates (North region is the largest one in Brazil). On the other hand, the smallest losses occur in the South, an opposite scenario of what can be observed in the northern region. In Southeast and Northeast, some utilities suffer with energy theft motivated by the large number of slums presented in states such as Rio de Janeiro, São Paulo, Bahia and Pernambuco. In Midwest and South, where the number of slums is smaller, the occurrence of frauds is more usual in the unit of metering of each consumer.

The amount of illegal connections in Brazil and the rates regarding them give us an idea of the magnitude of the problem and the degree of difficulty to compute such losses. According to Table II [3], which shows a list of the 15 utilities with the largest commercial losses rates,[1] the Centrais Elétricas do Pará (CELPA) leads the ranking with 24.4% of the distributed energy, followed by LIGHT company, located in Rio de Janeiro, where the losses reach 24.2% of the distributed energy. Finally, Centrais Elétricas de Rondônia (CERON) takes the third place with 22%. These three companies are located in North and Southeast regions, corroborating the data available in Table I. The main issue concerning commercial losses is related to the impact in the energy tariff, since this

[1]The commercial losses rates below 10% are tolerated by ANEEL. According to Millard and Emmerton [5], the commercial losses rates in Brazil were between 0.5% and 25.0% in 2007.

kind of loss ends up being divided among legal consumers registered in the electric utility at the time of tariff calculation. In a concession area such as of the LIGHT, for example, the tariff reduction would be of 18% if there was no irregular consumption [3].

A. Procedures to Combat Commercial Losses

The problem of commercial losses detection in distribution systems has been decisive. Theft and tampering of energy meters in order to modify the measurement of energy consumption are the main causes of commercial losses in electric utilities. Therefore, to calculate or even measure the amount of these tasks has been a difficult task. In most cases, it is almost impossible to know where they occur.

Aiming to reduce the rates concerning commercial losses, the electric utilities usually operate in the following preventing programs:

- Inspection Programs: they consist in verifying the integrity of the measurement system, to detect equipment failures, frauds and energy thefts, connection errors and other problems that may compromise the measurement of electric energy;
- Replacement of energy meters: it consists in the assessment of energy meters through field sampling, laboratory testing and analysis of the energy meters removed in the field. In addition, the replacement of energy meters with service life expired or possible technical failures is also performed;
- Regularization of illegal connections: especially in slums, through a regulatory program to reduce commercial losses;
- Implementation of trade policies: it consists in giving attention to the community about explanations, agreements and trainings in healthy energy consumption; and

One of the most traditional ways to combat commercial losses is to perform periodic inspections of consumers, which is not very advantageous, since such task has high costs to the electric utility. Additionally, the selection of consumers that must be inspected is an arduous task, even for experts. In the last decade, the electric utilities have invested much effort in a set of heuristic methods to automatic recognize illegal consumers by means of artificial intelligence-based techniques, which is the main focus if this paper.

B. Smart Grid and Its Relation to Commercial Losses

The problem of commercial losses may be minimized in the nearby future by means of smart metering resources. In addition to control consumption in real time, it is possible to collect more electrical information through sensors in energy meters, thus enabling a better understanding of the consumer behavior, and to transmit them with certain periodicity to the utilities. This can be seen as an advance with respect to the measurement process, but without employing the concept of machine learning for taking decisions autonomously. In other words, the smart meter does not prevent the fraud, but only provides information more quickly [2], [26], [27].

The computational intelligence aims to point out where is more likely to happen a fraud or irregularity, as well as what is its importance level through a priority criterion, followed by a field inspection. Thus, smart meters are extremely useful to improve the performance of the network, and also to reduce the commercial losses [26].

The concept of "Smart Grid", using smart meters, allows the integration of electrical equipment with data communication networks in a managed and automated system by the electric utility, making the energy to be supplied with safety, reliability and efficiency. Therefore, a network can enable [2], [26], [27]:

- Smart services integrated with consumers;
- The use of smart meters and the application of differentiated tariffs by the time or the seasonality;
- Improvement of power quality and reduction of technical and commercial losses;
- Management, monitoring and optimization of the energy system in real time;
- Surveillance and security;
- Integration of public services; and
- Broadband Internet.

The concept of Smart Grid has increased with the demand growth for automatic reading of energy meters. Beyond the aim of reducing frauds, thefts of energy and faulty measurements, several electric utilities have been concerned to better characterize the profile of consumers with irregularities, and also to correctly identify illegal connections [26], [27]. Therefore, minimizing commercial losses may guarantee investments in programs for power quality, and may allow a reduction of the price to the consumer.

III. Black Hole Algorithm

The Black Hole Algorithm is a population-based metaheuristic algorithm based on the black hole's gravitational force proposed by Hatamlou [24]. The candidate solutions (stars) are initialized at random positions onto the search space $\vec{x}_i \in \Re^n$ with $i = 1, 2, \ldots, m$, where n and m stand for the number of design variables and the number of stars, respectively. The objective function value of all stars are computed and the best star in the population, i.e., the one which holds the best objective function value, is selected to be the black hole.

All stars move toward the black hole due to its gravitational force absorbing everything that is around. As such, each star position is updated as follows:

$$\vec{x}_i^{(t+1)} = \vec{x}_i^t + \sigma\left(\vec{x}^* - \vec{x}_i^t\right), \tag{1}$$

where $\vec{x}_i^t$ is the location of the i-th star at iteration t, $\vec{x}^*$ is the location of the black hole in the search space, and $\sigma \sim U(0, 1)$. Notice that σ is different for each star and iteration. As the stars move toward the black hole, their positions keep changing and, consequently, it is assumed that their objective function value are getting better. If a star reaches a location with lower objective function value than the black hole, the star become the new black hole and all the other stars start to move toward the new black hole.

A sphere-shaped boundary known as "event horizon" surrounds the black hole swallowing everything that comes close.

Each star that crosses such boundary will be sucked by the black hole, and a new star borns randomly in the search space to start a new search. The radius r of the event horizon in BHA is calculated using the following formulation:

$$r = \frac{f^*}{\sum_{i=1}^{m} f_i}, \tag{2}$$

where f^* and f_i stands for the objective function values of the black hole and of the i-th star, respectively. When the distance between a star and the black hole is less than r, that star is collapsed and a new star is created. The next iteration takes place after all stars have been moved. Roughly speaking, the idea of BHA is to guarantee diversity in the population when creating new black holes, as well as to avoid stars getting trapped from local optima. This mechanism ends up contributing with both exploitation (local) and exploration (global) searches.

Unlike the standard BHA, in which the solutions are updated in the search space towards continuous-valued positions, in the proposed Binary Black Hole Algorithm (BBHA), the search space is modelled as an n-dimensional boolean lattice and the solutions are updated across the corners of a hypercube. In addition, as the problem is to select or not a given feature, a solution binary vector is employed, where 1 corresponds whether a feature will be selected to compose the new dataset, and 0 otherwise. In order to design this binary vector, we employed Equation (4), which can restrict the new solutions to only binary values:

$$S\left(x_{ij}^t\right) = \frac{1}{1 + e^{-x_{ij}^t}}, \tag{3}$$

$$x_{ij}^t = \begin{cases} 1 & \text{if } S(x_{ij}^t) > \gamma, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

in which $\gamma \sim U(0, 1)$ and x_{ij}^t stands for the j-th decision variable of the i-th star at iteration t. This approach is a slight variation of the one proposed by Nemati *et al.* [25]. Notice each variable to be optimized stands for one feature extracted from a given consumer (Section V-A). Therefore, each star models a binary-valued solution vector that indicates whether a feature will be selected or not to compose the final dataset.

IV. Case Study

Very often, the literature addresses commercial losses using an exhaustive search in spreadsheets in order to compare the historical consumption of thousands of consumers by hand. However, this procedure is time- and money-consuming, thus being interesting to make use of computational tools for accomplishing this task in a more efficient way.

The computational tools are often implemented using artificial intelligent in the context of machine learning research field. The theme addressed in this research does not only involve pattern recognition, but also optimization tasks, i.e., intelligent techniques are applied for optimization purposes considering feature selection purposes, aiming at characterizing the consumer profile to minimize commercial losses. Such optimization techniques can evidence the process of learning the behavior and characterization of potential consumers with irregularities. In this work, we validated the proposed technique for feature selection based on BHA against with PSO, HS, DE and GA in the context of theft characterization, as well as we provide an economic study based on the results in the Brazilian energy market.

The electric utilities usually look for methods more financially viable, being an affordable solution to employ softwares to support decision-making processes in face of thousands of consumers, pointing out those who may have some kind of error in their measurements. In other words, the software helps reducing the number of inspections, checking consumers suspected with irregularities and avoiding unnecessary inspections. Thus, it is possible to reduce costs with periodic random inspections, since this procedure will determine what may be the cause of the commercial loss, making sure the consumer is committing some kind of fraud or if the energy meter is reading the measurements correctly. Therefore, the utility can decide the best kind of providence that it should be taken to solve the problem quickly and effectively. Moreover, the utility will have a revenue recovery, because the irregular consumers will come back to be regular again, and they will return to properly pay for the consumed energy.

V. Methodology and Experimental Results

In this section, we present the experimental evaluation used to assess the effectiveness of BHA in the context of theft characterization (Section V-C), as well as we also discussed the impact with respect to the application of such techniques (Section V-D).

A. Datasets

A Brazilian electric utility has provided two private datasets, being one with 3,182 profiles of industrial consumers and the other with 4,952 profiles of commercial consumers, represented by eight features:

1) Demand Billed (DB): demand value of the active power considered for billing purposes, in kilowatts (kW);
2) Demand Contracted (DC): the value of the demand for continuous availability requested from the electric utility, which must be paid whether the electric power is used by the consumer or not, in kilowatts (kW);
3) Demand Measured or Maximum Demand (D_{max}): the maximum actual demand for active power, verified by measurement at fifteen-minute intervals during the billing period, in kilowatts (kW);
4) Reactive Energy (RE): energy that flows through the electric and magnetic fields of an AC system, in kilovolt-amperes reactive hours (kVArh);
5) Power Transformer (PT): the power transformer installed for the consumers, in kilovolt-amperes (kVA);
6) Power Factor (PF): the ratio between the consumed active and apparent power in a circuit. The PF indicates the efficiency of a power distribution system;
7) Installed Power (P_{inst}): the sum of the nominal power of all electrical equipment installed and ready to operate at the consumer unit, in kilowatts (kW);

8) Load Factor (LF): the ratio between the average demand ($D_{average}$) and maximum demand (D_{max}) of the consumer unit. The *LF* is an index that shows how the electric energy is used in a rational way.

At every 15 minutes, the electric utility recorded consumption data during one year. After that, such technical data was used to compute the aforementioned monthly features for both datasets. However, the company did not inform what kind of irregularity was verified in each consumer.

B. Experimental Setup

Basically, we employed the very same procedures of our preliminary work [2] concerning the experiments. Roughly speaking, in order to deal with the stochastic behavior of the optimization techniques, we ended up partitioning the dataset into *N* folds, in which two of them were used as training and validating sets, and the remaining folds were merged together to compose the test set. Therefore, such procedure is repeated over *N* times, and a statistical evaluation can be performed. The validating set is used to guide the optimization techniques, since the idea is to select the minimal subset of features that allows the best recognition rates over the validating set. Notice the test set does not participate from the learning features step. In regard to the recognition rate, we employed an accuracy measure proposed by Papa *et al.* [28], which considers unbalanced classes, such as the ones we faced in this work.

Although the reader can employ any supervised classification technique for the above procedure, in this paper we opted to employ the Optimum-Path Forest (OPF) classifier [28], [29], since it is a parameterless approach and it has obtained similar results to the ones achieved by some state-of-the-art pattern recognition techniques, being sometimes faster for training.

C. Theft Characterization

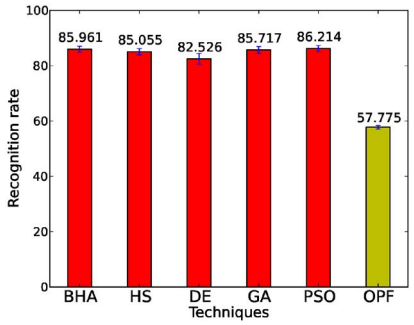
In this section, we present the results regarding to BHA, PSO, HS DE and GA techniques for theft characterization, i.e., we are interested to find out the most important set of features in order to identify possible illegal consumers. We employed $N = 5$ folds, a population of 30 candidate solutions (star/particle/harmony/chromosome) and 100 iterations for all techniques. The results presented in this section stand for the mean accuracy and standard deviation over 25 rounds using the methodology presented in Section V-B. Since the meta-heuristic techniques used in this work are non-deterministic, such approaches seem to be robust to avoid biased results. Table III presents the parameters used for each optimization technique.[2]

The exploration and exploitation of the metaheuristic algorithms is controlled by user parameters. PSO uses c_1 and c_2 for the pace range control, which guides the particles toward their best local solution as well as to the best global solution of the swarm, respectively. Additionally, the amount of velocity that is going to be used to update the value of each possible solution for the next time step is controlled by the inertia weight w.[3] In regard to HS, *HMCR* stands for the Harmony Memory Considering Rate, which controls the amount of information extracted from the previously used values to compose new solutions. In regard to DE, f is called differential weight, which scales the influence of the set of pairs of solutions selected to calculate the mutation value, and c_r stands for crossover probability. Finally, GA's parameter p_m denotes the mutation probability. Figs. 1a and 1b depict the recognition rates over the commercial and industrial datasets, respectively. The "yellow" bar stands for standard OPF recognition rate, i.e., without feature selection.

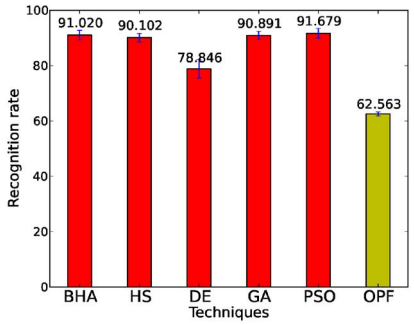
Observing Figs. 1a and 1b, we can note a great improvement with respect to standard results (i.e., naïve OPF), being all techniques similar to each other if we consider their standard deviation. Such experiment demonstrated BHA is able to achieve results similar to those obtained by well-accepted meta-heuristic techniques in the literature, such as PSO, HS, DE and GA. Another experiment evaluated the convergence rates of each optimization technique, as displayed in Figs. 2a and 2b: it is clear PSO and GA have converged faster, but it did not reflect on the final classification results displayed in Fig. 1. Although naïve HS may be considered one of the fastest approaches, it often does not benefit from reasonable convergence rates, since it updates only one agent (harmony)

TABLE III
PARAMETERS EMPLOYED FOR EACH OPTIMIZATION ALGORITHM

Algorithm	Parameters
PSO	$c_1 = c_2 = 2.0$ and $w \in [0.4, 0.9]$
HS	$HMCR = 0.9$
DE	$f = 0.5$ and $c_r = 0.1$
GA	$p_m = 0.1$
BHA	-



(a)



(b)

Fig. 1. Mean recognition rates over (a) commercial and (b) industrial datasets.

[2]The parameters have been empirically chosen.

[3]Parameter w has been dynamically adjusted within in interval $[0.4, 0.9]$.

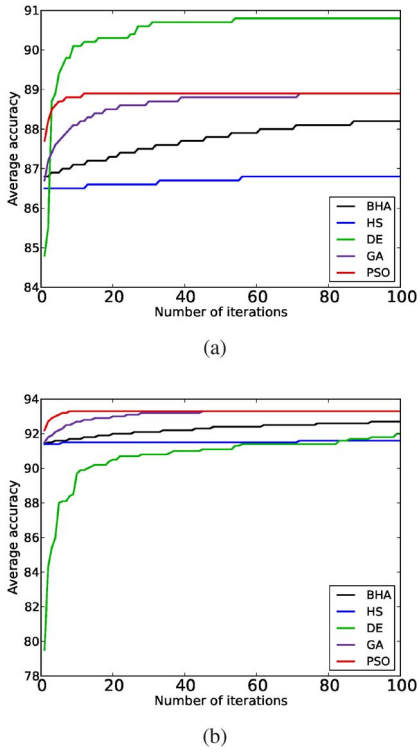


Fig. 2. Convergence rates over (a) commercial and (b) industrial datasets.

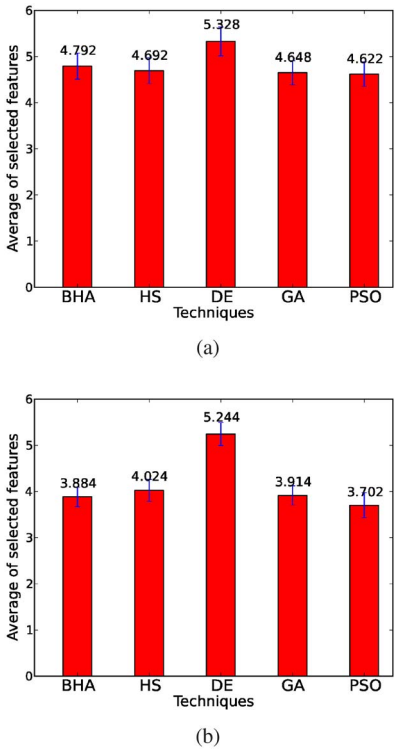


Fig. 3. Average number of selected features over (a) commercial and (b) industrial datasets.

at each iteration, while swarm-based techniques usually update all agents.

The average number of selected features for each optimization technique is shown in Figs. 3a and 3b concerning commercial and industrial profiles, respectively. Such experiment attempts to show that only 58.75% (on average) of the features really matter for illegal consumer recognition in case of commercial profiles, and 50% (on average) considering industrial dataset. We can observe the smallest number of features have been selected by PSO and BHA in this latter dataset.

Figs. 4a and 4b present the computational load (ms) for commercial and industrial datasets, respectively.[4] Two groups of techniques can be observed here: HS and swarm and genetic-based ones, which are composed by BHA, DE, GA and PSO. The latter approaches usually evaluate the entire swarm in order to update all agents' position, while HS only updates one agent at each iteration. Therefore, if we consider the second group, we notice BHA has been the fastest approach. Another interesting skill of BHA is related to its absence of parameters, which is very interesting to avoid meta-optimization, turning the technique user-friendly and less prone to configuration errors.

Additionally, we have evaluated results using the Wilcoxon statistical test [30], as displayed in Table IV. The Wilcoxon test evaluates each pair of techniques in order to check whether they are similar to each other or not. The symbol '$\neq$' denotes there exists difference between the methods, and

[4]The experiments were executed on a computer with a Pentium Intel Core *i*7® 1.73Ghz processor, 6 GB of memory RAM and Linux Ubuntu Desktop LTS 13.04 as the operational system.

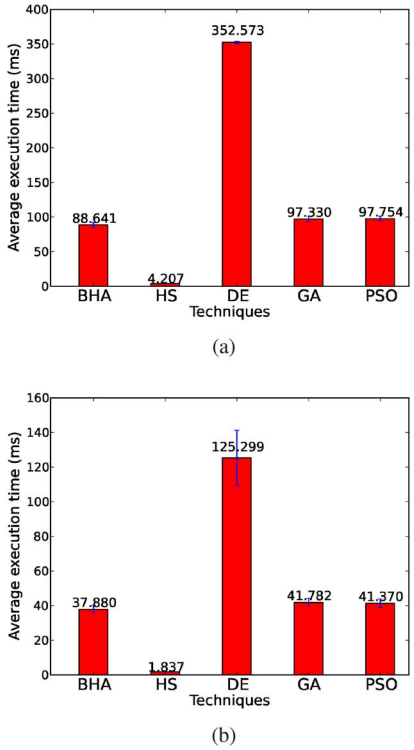


Fig. 4. Mean execution time (ms) over (a) commercial and (b) industrial datasets.

the symbol '$=$' represents the techniques are similar each other. Additionally, if the p-value (parenthesis) is less than the desired significance level, the techniques are considered different to each other. In this paper, we adopted 0.05 (5%) of

TABLE IV
WILCOXON SIGNED RANK TEST CONSIDERING 5% OF SIGNIFICANCE

Dataset	BHA/HS	BHA/DE	BHA/GA	BHA/PSO
Commercial	$\neq$ (0.0000)	$\neq$ (0.0000)	$\neq$ (0.0422)	$\neq$ (0.0054)
Industral	$\neq$ (0.0000)	$\neq$ (0.0000)	= (0.3395)	$\neq$ (0.0000)

TABLE V
MEAN ACCURACY RATES PER CLASS WITHOUT FEATURE SELECTION

Dataset	Regular Consumer	Irregular Consumer
Commercial	95.01%	58.87%
Industrial	94.86%	64.14%

TABLE VI
MEAN ACCURACY RATES PER CLASS WITH FEATURE SELECTION

Dataset	Regular Consumer	Irregular Consumer
Commercial	97.54%	64.81%
Industrial	98.94%	83.76%

significance. According to Table IV, PSO was the most accurate technique with respect to commercial dataset, followed by BHA. The same result can be evidenced for industrial dataset. However, the accuracy rates of PSO and BHA are very close to each other, being BHA faster and parameter-free, which makes it suitable for feature selection purposes.

D. Impacts

In this section, two classes are considered for the experiment: (i) the "Regular Consumer", which represents consumers who are under regular conditions, and (ii) the "Irregular Consumer", which stands for potential consumers with irregularities. It is important to highlight the electric utility did not provide any further details about irregularities presented in each consumer, but they were previously confirmed by the technical staff of the electric utility. In this section, the mean accuracy was computed to verify a comparison between different classes, i.e., the also computed the recognition rates for each type of consumer (regular or irregular). Table V presents the accuracy rates per class without feature selection.

Observing Table V, we can note that 58.87% of commercial consumers and 64.14% of the industrial consumers who had some kind of irregularity were identified correctly. Therefore, the regular consumers have higher recognition rates, probably because the dataset is biased on such class, i.e., we have much more regular consumers than irregularities in both datasets.

Table VI presents the accuracy rates per class considering the feature selection by means of BHA. We can observe that 64.81% of commercial consumers and 83.76% of industrial consumers who had some kind of irregularity were identified correctly. Thus, the accuracy rates increased approximately 20% for industrial dataset and 6% for commercial dataset when compared with the vanilla results (without feature selection). Therefore, the results were quite optimistic.

Table VII presents the selected features considering both datasets. Notice these features are extracted from a single execution of the algorithms, and may not reflect the final subset of features, since the experiments average the number of them.

TABLE VII
SELECTED FEATURES CONSIDERING BHA

Features for Commercial	Features for Industrial
DC, D_{max}, PF, P_{inst}, LF	DC, D_{max}, PF, P_{inst}, LF

The same set of features have been chosen for both datasets, which highlights their importance.

VI. CONCLUSION

We presented here the context of commercial losses (non-technical losses) in Brazil, and discussed how this issue is recent. Note that in less developed regions, where the socio-economic aspects, such as education, income distribution and violence, among others, are very precarious, the rates of commercial losses are extremely high. The fraud and theft of energy are attitudes of many consumers unable to pay for the consumed energy, or even malicious behaviour in order to save money. Hence, there is a need for several ways to combat these commercial losses in order to not compromise the power system, thus generating electric energy with quality, as well as to possibly reduce the final energy price to consumers.

The development of intelligent computational tools has been widely pursued to contribute to the reduction of commercial losses, since these methods can be easily employed in smart grids, which will be deployed worldwide. However, these tools only assist in decision-making processes to indicate a potential consumer with irregularities, which is checked after an inspection conducted by the technical staff of the electric utility, i.e., they do not confirm whether the consumer is fraudster or not. The most works address only the identification or detection of commercial losses. In this paper, we are concerned about characterizing the profile of possible irregular consumers, i.e., we want to determine the most relevant features considering the context of the problem.

We also presented a case study to demonstrate the usefulness of the methodology described in this work which was conducted by means of meta-heuristic techniques and OPF classifier, as well as we have introduced BHA for feature selection purposes in the context of commercial losses in power systems.

REFERENCES

[1] M. E. de Oliveira, D. F. A. Boson, and A. Padilha-Feltrin, "A statistical analysis of loss factor to determine the energy losses," in *Proc. IEEE/PES Transm. Distrib. Conf. Expo. Latin America*, Bogotá, Colombia, 2008, pp. 1–6.

[2] D. Rodrigues, C. C. O. Ramos, A. N. de Souza, and J. P. Papa, "Black hole algorithm for non-technical losses characterization," in *Proc. IEEE 6th Latin Amer. Symp. Circuits Syst. (LASCAS)*, Montevideo, MN, USA, 2015, pp. 1–4.

[3] ANEEL, "Irregular power consumption generates loss of R$ 8,1 billion per year," Clic Energia, 2011.

[4] V. Paruchuri and S. Dubey, "An approach to determine non-technical energy losses in India," in *Proc. 14th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2012, pp. 111–115.

[5] R. Millard and M. Emmerton, "Non-technical losses—How do other countries tackle the problem?" in *AMEU Proc.*, Cape Town, South Africa, 2009, pp. 67–81.

[6] EPE, "Statistical yearbook of electricity 2014–2013 baseline year," Energy Res. Company, Rio de Janeiro, Brazil, Tech. Rep., 2014.

[7] S.-C. Huang, Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2959–2966, Aug. 2013.
[8] J. A. Porras, H. Rivera, F. D. Giraldo, and B. S. Acosta, "Identification of non-technical electricity losses in power distribution systems by applying techniques of information analysis and visualization," *IEEE Latin Amer. Trans.*, vol. 13, no. 3, pp. 659–664, Mar. 2015.
[9] R. Alves, P. Casanova, E. Quirogas, O. Ravelo, and W. Gimenez, "Reduction of non-technical losses by modernization and updating of measurement systems," in *Proc. IEEE/PES Transm. Distrib. Conf. Expo. Latin America*, Caracas, Venezuela, 2006, pp. 1–5.
[10] C. C. O. Ramos, A. N. Souza, J. P. Papa, and A. X. Falcão, "Fast non-technical losses identification through optimum-path forest," in *Proc. 15th Int. Conf. Intell. Syst. Appl. Power Syst. (ISAP)*, Curitiba, Brazil, 2009, pp. 1–5.
[11] J. Nagi, A. M. Mohammad, K. S. Yap, S. K. Tiong, and S. K. Ahmed, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.
[12] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.
[13] C. C. O. Ramos, A. N. de Souza, J. P. Papa, and A. X. Falcão, "A new approach for nontechnical losses detection based on optimum-path forest," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, Feb. 2011.
[14] I. Monedero, F. Biscarri, C. León, J. Biscarri, and R. Millán, "MIDAS: Detection of non-technical losses in electrical consumption using neural networks and statistical techniques," in *Proc. Int. Conf. Comput. Sci. Appl.*, vol. 3984. Glasgow, U.K., 2006, pp. 725–734.
[15] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and F. Nagi, "Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system," *IEEE Trans. Power Del.*, vol. 26, no. 2, pp. 1284–1285, Apr. 2011.
[16] A. H. Nizar, J. H. Zhao, and Z. Y. Dong, "Customer information system data pre-processing with feature selection techniques for non-technical losses prediction in an electricity market," in *Proc. Int. Conf. Power Syst. Technol.*, Chongqing, China, 2006, pp. 1–7.
[17] C. C. O. Ramos, A. N. Souza, G. Chiachia, A. X. Falcão, and J. P. Papa, "A novel algorithm for feature selection using harmony search and its application for non-technical losses detection," *Comput. Elect. Eng.*, vol. 37, no. 6, pp. 886–894, 2011.
[18] C. C. O. Ramos, A. N. de Souza, A. X. Falcão, and J. P. Papa, "New insights on nontechnical losses characterization through evolutionary-based feature selection," *IEEE Trans. Power Del.*, vol. 27, no. 1, pp. 140–146, Jan. 2012.
[19] J. F. Kennedy and R. C. Eberhart, *Swarm Intelligence*. San Francisco, CA, USA: M. Kaufman, 2001.
[20] R. Storn and K. Price, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Glob. Optim.*, vol. 11, no. 4, pp. 341–359, 1997.
[21] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.
[22] Z. W. Geem, *Music-Inspired Harmony Search Algorithm: Theory and Applications*, 1st ed. Heidelberg, Germany: Springer-Verlag, 2009.
[23] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, 1997.
[24] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," *Inf. Sci.*, vol. 222, pp. 175–184, Feb. 2013.
[25] M. Nemati, H. Momeni, and N. Bazrkar, "Article: Binary black holes algorithm," *Int. J. Comput. Appl.*, vol. 79, no. 6, pp. 36–42, 2013.
[26] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid—The new and improved power grid: A survey," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 944–980, Dec. 2012.
[27] M. R. Guarracino, A. Irpino, N. Radziukyniene, and R. Verde, "Supervised classification of distributed data streams for smart grids," *Energy Syst.*, vol. 3, no. 1, pp. 95–108, 2012.
[28] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki, "Supervised pattern classification based on optimum-path forest," *Int. J. Imag. Syst. Technol.*, vol. 19, no. 2, pp. 120–131, 2009.
[29] J. P. Papa, A. X. Falcão, V. H. C. de Albuquerque, and J. M. R. S. Tavares, "Efficient supervised optimum-path forest classification for large datasets," *Pattern Recognit.*, vol. 45, no. 1, pp. 512–520, 2012.
[30] T. Harris and J. W. Hardin, "Exact Wilcoxon signed-rank and Wilcoxon Mann–Whitney ranksum tests," *Stata J.*, vol. 13, no. 2, pp. 337–343, 2013.

Caio C. O. Ramos received the B.Sc. and M.Sc. degrees from São Paulo State University, SP, Brazil, in 2006 and 2010, respectively, and the Ph.D. degree from the University of São Paulo, SP, in 2014, all in electrical engineering. He is currently a Post-Doctorate Researcher with São Paulo State University. His interests include intelligent systems, power quality, and nontechnical losses in electrical systems.

Douglas Rodrigues received the degree in business management and informatics from the Faculdade de Tecnologia de Botucatu, SP, Brazil, in 2009. He is currently pursuing the M.Sc. degree in computer science with São Paulo State University. His research interests include machine learning and pattern recognition.

André. N. de Souza received the B.Sc. degree from Mackenzie University, SP, Brazil, in 1991, and the M.Sc. and Ph.D. degrees from the Polytechnic School, University of São Paulo, SP, in 1995 and 1999, respectively, all in electrical engineering. He has been an Associate Professor with the Electrical Engineering Department, São Paulo State University, since 2005. His interests include intelligent systems, transformers, fraud detection in electrical systems, atmospheric discharges, and power quality.

João P. Papa received the B.Sc. degree in information systems from São Paulo State University, SP, Brazil, the M.Sc. degree in computer science from the Federal University of São Carlos, SP, in 2005, and the Ph.D. degree in computer science from the University of Campinas, SP, in 2008. He has been a Professor with the Computer Science Department, São Paulo State University, since 2009. His research interests include machine learning, pattern recognition, and image processing.