Unsupervised Classification for Non-Technical Loss Detection

George M. Messinis
Nikos D. Hatziargyriou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
[gmessinis, nh]@power.ece.ntua.gr

Abstract— **Electricity theft has been a major issue for DSOs for years. However, despite efforts to detect it and the application of legal deterrents, the phenomenon insists. In this paper, unsupervised data mining NTL detection techniques are tested on a smart meter data set provided by the Greek DSO, HEDNO and a publically available smart meter data set. Frauds are simulated and the Twitter breakout detection library is used for extracting features which will be combined with typical features already found in literature. Then, unsupervised classifiers such as rule systems, multi-variate Gaussian distribution (MGD), local outlier factor (LOF), k-means, fuzzy c-means, DBSCAN and SOM are demonstrated. Finally, a number of metrics are calculated for each data set and classifier such as accuracy, detection rate, precision, false positive rate and F_1 score. In addition, the amount of stolen energy detected and detection delay are proposed as metrics for NTL detection system studies.**

Index Terms-- **Electricity theft, non-technical losses, outlier detection, smart meters, unsupervised learning.**

I. Introduction

Electricity theft affects the financial income of energy suppliers and operators. Since in most cases the related costs are transferred together with the technical losses to benign consumers, increasing their bills, it is an anti-social phenomenon that needs to be combated. However, despite the efforts by the DSOs to detect it and the application of legal deterrents, the phenomenon insists. In Greece, non-technical losses were estimated close to 2% in 2014 which translates to a conservative estimation of €150 M/year including next to the value of energy, the corresponding network charges taxes, levies, etc.

A number of NTL (Non-Technical Loss) detection methods have been presented in literature, mainly after 2010 [1]. These methods may be categorized as data oriented, network oriented and hybrids. Network oriented methods make use of network resources such as grid topology or RTU (Remote Technical Unit) data [2]- [4]. Data oriented methods are network agnostic and only require consumer related information (such as energy consumption measurements, demographic data etc) [5]-[7]. Hybrid methods are a combination of the previous two [8], [9]. Data oriented methods are divided in supervised and unsupervised methods. Unsupervised methods have received less attention in literature. They do not require fully labeled data sets and thus can be applied easier in real life applications. Unsupervised methods are also used, in case a large number of negative (benign) samples are available, while positive samples (fraud) are scarce or not available at all. Another advantage of unsupervised methods is their resilience to zero day attacks. The aforementioned characteristics of unsupervised methods make them more suitable for fraud detection and are thus the focus of this paper.

Another important focus of this paper concerns feature extraction from available time series which best describe the fraud phenomenon. Although the active energy time series itself may be used for such purposes, extracting a small number of highly representative features is much more efficient. A number of unsupervised NTL detection and relative features reported in literature are shortly described next. Among them the Twitter breakout detection package is used for the first time for fraud detection [20].

II. Related Work

The SOM (Self Organizing Map) has been used in [10] for classifying weekly consumption vectors of each consumer as fraud or not. Clustering can also be used as an unsupervised classification tool. Fuzzy clustering (fuzzy c-means) has been used in [11], where features such as average, maximum and standard deviation of consumption for the last 6 months are calculated. Another clustering approach is presented in [12]. In this case the authors use Principal Component Analysis (PCA) for extracting features and the Density Based Spatial Clustering of Applications with Noise (DBSCAN) for clustering data. Expert systems (rules) have been used in [13] together with text mining and neural networks, while fuzzy rules where used in [5] to improve the performance of a supervised SVM classifier. Statistical process control, like the Shewhart individuals (XmR) control chart proposed in [14] may also be used for detecting frauds. Such methods can also be used in real time and do not require extracting features. Regression models have been used for forecasting a time series in [15]. The main concept behind using forecast models for NTL detection is the following: given a set of non-malicious measurements a model is trained to forecast the next point of the time series. The measured and

forecasted values are then compared. LOF (local outlier factor) is a density based outlier detection technique used as an indicator for fraud in [16].

Unfortunately, almost all of the presented algorithms are applied on different data sets making difficult to compare the various approaches. Additionally, fraud samples were not available for both data sets since the objective of the two pilots was not initially related to non-technical loss detection. Thus, frauds are simulated using different models and parameter values. Finally, a common definition and use of performance metrics is rare among literature and sometimes important metrics tied with the nature of the fraud detection problem (class imbalance, base rate fallacy) are neglected. The next chapter is focused exactly on these issues.

III. Modeling NTL and Measuring Performance

A. Data Set Description

Two different data sets will be used for demonstrating feature extraction and unsupervised methods for NTL detection. The first data set is provided by the Hellenic Distribution Network Operator (HEDNO), the Greek DSO, and contains smart metering measurements from 500 commercial consumers (12 months). The active energy consumption is recorded with a time resolution of 15 minutes. The second data set is publicly available since 2012 and is provided by the Commission for Energy Regulation (CER) of Ireland, [17]. It includes half hourly active energy consumption of about 5000 residential and commercial consumers during 2009 and 2010. Both data sets are assumed to be free of NTL and thus fraud will be simulated as presented in the next paragraph.

The R programming language is used for all purposes of this work. R is especially designed for data analytics and data mining processes making it the perfect candidate for smart meter time series analytics like fraud detection. In addition, R is under the GNU General Public License making it easy to reuse a vast number of libraries. R runs on all UNIX-like operating systems. All simulations are performed on a Windows 7, Intel Core i5-430M @ 2.27 GHz processor with 6GB of RAM machine.

B. Modelling Non-Technical Losses

The absence of smart metering as well as difficulties in detecting frauds inevitably lead to a shortage of fraudulent samples. Such samples are of utmost importance for verifying the operation of any NTL detector. In addition, at least as far as the CER data set is concerned, no frauds are expected in the data set, since consumers themselves agreed to take part in the research. In such cases fraud must be simulated. Simulated fraudulent samples may not be original, but given an accurate electricity theft model, it is possible to control the training and test sets and perform sensitivity analysis on various parameters that influence electricity theft. This enables testing the NTL detector under various circumstances.

A number of electricity fraud models have been proposed in literature, including those described in [8] and [18]. A more simplistic model (Fig. 1) is employed in this case, in order to capture the effects of electricity fraud on yearly consumption curves. Thus, assuming a NTL detector which receives as input the active energy consumption for a duration of 365 days per consumer there are two main parameters of interest:

- Fraud Intensity: This parameter expresses the amount of energy stolen by the malicious user. Active energy consumption measurements are multiplied by this factor in order to simulate fraud. Fraud intensity is assumed constant throughout the year, although it may vary per consumer.
- Fraud Start: This parameter represents the day of year fraud is initiated by a consumer. If fraud is initiated early in the period of interest when data is available (or was present even before that) no significant changes will be visible in the consumption time series. Additionally, if fraud is initiated late in the period of interest a visible change in the time series will not be enough to infer fraud.

Fig. 1 presents the fraud model employed for simulating NTL in the previously described data sets. The data availability period is defined as the time period for which consumption measurements with specific requirements (for example at least daily energy usage measurement) are available. For the purposes of this research, the data availability period is equal to 365 days. In addition, an attempted fraud is expected to impact the consumption time series by inserting an anomaly at the point fraud is initiated (fraud start). Detecting this anomaly is a cornerstone of the proposed NTL detection scheme.

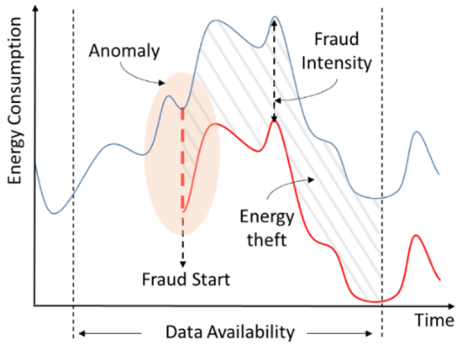

Figure 1: Electricity theft model. Blue line represents a benign consumption profile, while the red line represents the same profile with fraud.

C. Performance Metrics

A number of performance metrics for classification problems can be found in literature. Most of them are calculated from the confusion matrix where TP are true positives, TN are true negatives, FP are false positives and FN are false negatives. Some of the most popular metrics are:

$$False\ Positve\ Rate\ (FPR) = \frac{FP}{FP + TN}$$

$$Detection\ Rate\ (DR) = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F_1\ score = \frac{2TP}{2TP + FP + FN}$$

Although these metrics may provide a good view of the classifiers' performance, they neglect a number of important issues emerging from the fraud detection problem nature. Three more metrics are thus defined:

- Bayesian detection rate (BDR): Base rate fallacy [19] has not yet been extensively covered in NTL detection literature. This is why BDR only appears in [8], although it reveals a lot on how an NTL detector would work in real life. BDR is calculated as follows:

$$BDR = \frac{P(I) \cdot DR}{P(I) \cdot DR + P(\neg I) \cdot FPR}$$

where:
$P(I)$: the probability of fraud
$P(\neg I)$: the probability of no fraud

The probability of fraud is an external parameter that cannot be influenced and expresses the frequency of the phenomenon. This parameter usually receives small values (1%-5%), since fraud is typically not frequent. Thus, in order to achieve a high value for BDR (i.e. minimize false alarms) an extremely low value for FPR must be achieved even if DR is high.

- Fraud energy detected: The percentage of energy theft detected calculated by summing up electricity theft of correctly classified consumers and dividing by the total amount of electricity theft. This metric indicates if large chunks of electricity theft are detected even though small numbers of frauds are found.
- Detection delay: The deviation between the original day fraud was initiated by a consumer (fraud start) and the day fraud was initially detected by the NTL detector for correctly classified consumers. This metric provides insight on how accurate the NTL detector is, when tracking anomalies in a consumption time series. Additionally, the day fraud was initiated, as defined by the NTL detector, can be used for calculating the total energy stolen, and thus must be accurate as possible.

IV. Extracting Features from Data

Feature extraction is quite typical in many classification processes. The need to manage high dimensional time series data for large numbers of consumers makes feature extraction of utmost importance. Although the time series itself could be used when designing a NTL detection system, there are various drawbacks in doing so, such as the large data volume available in time series, the personal information time series carry, visualization difficulties when assessing large numbers of time series etc. A lot of work has been conducted on time series feature extraction and selection [14], but as far as electricity fraud detection is concerned, there are no specific guidelines. Taking into account the above considerations the authors dedicate this section to feature extraction for detecting NTL.

A. The Twitter breakout detection package

Twitter announced in 2014 the Breakout Detection package, an open source R package for detecting breakouts in cloud data [20]. According to this, a breakout is defined as a mean shift (a sudden jump in the time series) or as a ramp-up (a gradual increase/decrease from one steady state to another) in the time series. The concept of "mean shift" strongly applies to most cases of energy fraud too, assuming that the malicious user is not committing any kind of "smart" attack. The existence of a sudden drop in the energy consumption time series is generally considered as an indicator of fraud. Breakouts could be a result of non-malicious behavior too (e.g. replacement of energy consuming device, change in number of residents). Such cases would probably result to false positives, that can be avoided by combining a classifier's output with a set of relevant rules defined by experts.

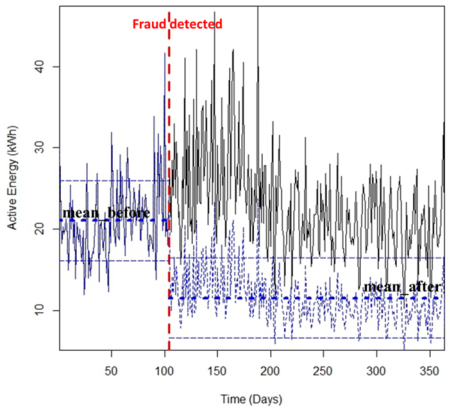

Figure 2: The Twitter BreakoutDetection package applied on a CER data set consumer. The black line is the original timeseries while the blue is the same consumer with simulated fraud.

When reviewing an active energy consumption time series however, sudden drops of consumption frequently appear, especially when dealing with 15 minute or hourly measurements. Even aggregating consumption to daily measurements will not solve the problem of multiple sudden jumps in consumption. A commercial consumer for example would still experience sudden drops at least each Sunday. Using a weekly or monthly energy consumption time series (or smoothing the time series) on the other hand may result in hiding all sudden drops including those that are results of fraud.

The Breakout Detection toolbox promises sufficient and fast breakout detection in the presence of anomalies and measurement noise. The algorithm used is described in [21] and employs energy statistics for detecting breakouts. An example of applying the Breakout Detection toolbox to a specific consumer of the CER data set is given in Fig. 2. Fraud is initiated in day 104 of the year and has an intensity of 0.5. The black line represents the consumer when no fraud is simulated, while the blue line represents the same consumer after fraud is simulated. In this case the Twitter Breakout Detection correctly identifies the abrupt change in consumption caused by fraud. In addition, one can easily spot the difference of mean consumption before and after the detected event which is close to 50%.

B. Feature definitions

The Breakout Detection library is used for detecting events that might indicate frauds in the time series. Application of the library to a non-malicious user time series however, might erroneously result in a possible fraud event of lower significance. The statistical significance of the breakout is provided by the package itself, but more simplistic features (comprehensible by humans) are chosen in this paper instead:

1. Normalized change in mean: In cases of fraud, a change in the mean of a consumption time series E should be observed after a breakout is detected at time τ. The difference between the mean consumption before and after the breakout is computed. This difference is then normalized by dividing with the mean yearly consumption.

$$\Delta\bar{E} = \frac{mean(E[1,\tau]) - mean(E[\tau,end])}{mean(E)}$$

2. Change in standard deviation: In cases fraud is modelled as in Chapter III, the standard deviation of the consumption after a fraud event is expected to be lower than the one before.

$$\Delta\sigma = \frac{sd(E[1,\tau]) - sd(E[\tau,end])}{sd(E[1,\tau])}$$

Apart from the above, two more features are defined in order to efficiently separate malicious from non-malicious users. These features are not as comprehensive as the previous two and were chosen after randomly testing a number of features present in literature:

3. Im: Imaginary part of the normalized time series Fourier transformation
4. Cos: Second component of the normalized yearly time series discrete cosine transformation

Calculation of the aforementioned features requires smart metering data (at least daily consumption) and can either be realized after gathering a significant amount of data or in real time, assuming a rolling window of the same length.

V. Unsupervised NTL Detection Techniques

In this chapter a number of different unsupervised NTL detection methods will be presented. All methods require at least a year of historical consumption data with a time resolution of 1 hour or less. A number of 3639 residential consumers from the CER data set are used for demonstrating each method before comparing them in Chapter IV. The probability of fraud *P(I)* (percentage of consumers committing fraud) is set to 5% in order to realistically evaluate BDR. This figure is chosen after taking into account HEDNO's experience on detecting NTL and in order to demonstrate realistic scenarios of fraud. Larger numbers would lead to higher BDR values although excessively increasing this percentage would probably lead to the deterioration of most anomaly detection methods (since NTL does not constitute an anomaly in this case). Fraud start is chosen randomly between day 40 and day 290, while fraud intensity follows a normal distribution with a mean of 0.4 and standard deviation of 0.08.

A. Expert System

An Expert System consists of a set of rules. These rules may be inferred from the data set (supervised learning) or may be defined by experts according to their experience in detecting NTL. The main issue with rules is that they are not easily devised for features difficult to comprehend. In this case, an NTL detection expert may not be able to define the impact of fraud on features such as *Im* and *Cos* described in the previous chapter. On the other hand, it is much easier to define rules for $\Delta\bar{E}$ and $\Delta\sigma$ and this is why only these two features are used. The rules applied are presented below. $\Delta\bar{E}_{min}$ and $\Delta\sigma_{min}$ are parameters that influence the NTL detector's sensitivity. Fig. 3-a reveals the influence of these two parameters on the detector's F_1 score and suggests that $0.55 < \Delta\bar{E}_{min} < 0.85$ and $\Delta\sigma_{min} < 0.25$ for obtaining fair results.

$$Rule\ 1: If\ \Delta\bar{E} > \Delta\bar{E}_{min}\ then\ FRAUD$$

$$Rule\ 2: If\ \Delta\bar{E} < 0\ then\ NOT\_FRAUD$$

$$Rule\ 3: If\ \Delta\sigma > \Delta\sigma_{min}\ then\ FRAUD$$

$$Rule\ 4: If\ \Delta\sigma < 0\ then\ NOT\_FRAUD$$

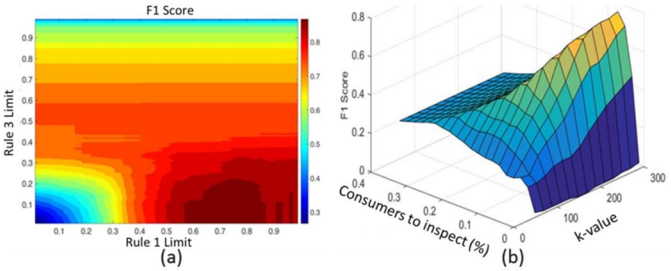

Figure 3: Parameter impact analysis: (a) Expert system F_1 score with respect to rule limits and (b) LOF F_1 score with respect to the percentage of inspected consumers and k-value

B. Local Outlier Factor

Local Outlier Factor (LOF) is a density based anomaly detection method calculating the local density of a data point with respect to the local density of its neighbors. LOF is computed for all consumers who are then ranked and a top percentage of them are marked as frauds. The number of consumers marked as frauds depends on the number of consumers a utility can inspect. The main drawback of LOF is that it detects outliers which are not necessarily due to fraud. Large negative values of $\Delta\bar{E}$ (abrupt increase of consumption) may be marked as fraud by LOF for example. This is why LOF is combined with Rules 2 and 4 presented above. In addition, the performance of the LOF detector depends on the k-value which, simply put, represents the maximum number of consumers that could be considered as outliers. Fig. 3-b suggests a k-value between 150 and 300 (which was expected since the total number of consumers equals 3639). In addition, the percentage of consumers to inspect should vary along 5-15% for achieving high F_1 score values. Both of these values can be derived just from an estimation of the probability of fraud in the data set (no labeled samples are required).

C. Multivariate Gaussian Distribution

In this case the data are modeled using a Multivariate Gaussian Distribution (MGD). The probability density function is calculated as:

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

where:

x: feature vector, μ: mean vector, Σ: covariance matrix, n: size of feature vector

After calculating the probability of each sample consumers are listed with ascending probability. Next Rules 2 and 4 may be applied to exclude outliers which are not frauds from the list.

Finally, a top percentage of the list is chosen for inspection (i.e. marked as fraud). The percentage of inspected consumers is expected to have high impact on the results as shown for the previous methods. Choosing to inspect 5% of the total number of consumers and after applying Rules 2 and 4 results to F1 score=88.12%, DR=92.77% and FPR=0.93% making the MGD approach a promising applicant.

D. Clustering

Various clustering algorithms have been used in unsupervised classification tasks. The main issue with such approaches is the number of clusters and how to interpret them. The most straightforward approach would be to define 2 clusters where the smaller one represents frauds. This assumes that frauds share the same characteristics on the feature space. On the other hand, one could argue that NTL is expressed in various ways and thus cannot be grouped under a single cluster. Choosing a number of clusters larger than 2 could solve this problem. In this case the cluster centers (average cluster time series) are evaluated and the one with highest negative slope is considered to be the fraud cluster. This heuristic approach may not always work though. The number of clusters must be carefully chosen and their centers should be evaluated by an expert for deciding which of them represent fraud.

The k-means, fuzzy c-means (FCM) and DBSCAN algorithms are considered in this work. The number of clusters equals 2 for k-means and FCM. Running k-means on the data set results to F_1 score=81.3%, FPR=1.67% and DR=90.55%. As far as FCM is concerned the membership value of each data point to the smallest cluster is used for ranking the consumers and a top percentage of them are then marked as frauds. A sensitivity analysis on the influence of the fuzzification factor as well as the percentage of consumers to be inspected on F_1 score reveals that setting the fuzzification factor in the range of 1.5-2 and the consumers to be inspected close to 5% will produce satisfactory F1 score values.

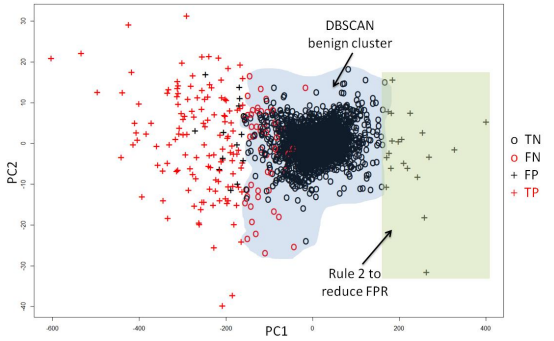

Figure 4: Application of PCA-DBSCAN-Rule 2 combination

DBSCAN is used after applying PCA for producing a 2D feature space. The use of PCA is not expected to dramatically increase the classifier's performance (since all 4 features are highly descriptive), it does however allow for a visual inspection of the algorithm results (Fig. 4). DBSCAN is tuned to produce two clusters, one of which represents outliers. The MinPts (minimum number of points required to form a dense region) parameter is set to 200 (approximately the number of frauds expected in the data set). The value of ε is chosen by determining the knee of the k-distance graph (where k equals to MinPts). In order to enhance the PCA-DBSCAN performance, Rule 2 is also applied thus reducing FPs. In this case F1 score=75.4%, DR=71.7% and FPR=0.9%.

E. Self Organizing Map (SOM)

SOM is essentially a type of neural network frequently used for dimensionality reduction. It is not a classification method itself, but can be used as core part of a classifier. SOM has the ability to produce fine 2D visualizations of high dimensional data. Its output may either be comprehended by an expert (via inspecting the visualization) or fed to a simple clustering algorithm, as described in the previous paragraph. In this work k-means (number of clusters equals 2, although it could be higher) is combined with SOM for detecting frauds. The SOM grid is presented in Fig. 5. The lattice size is chosen heuristically (total number of 400 nodes) and depends on the size of the data set. All samples belonging to the smaller cluster produced by k-means are marked as frauds. In addition, k-means is executed 100 times and the best clustering is then chosen. F1 score for this particular data set is equal to 89.66%, while FPR=0.8% making the SOM/k-means combination a promising applicant.

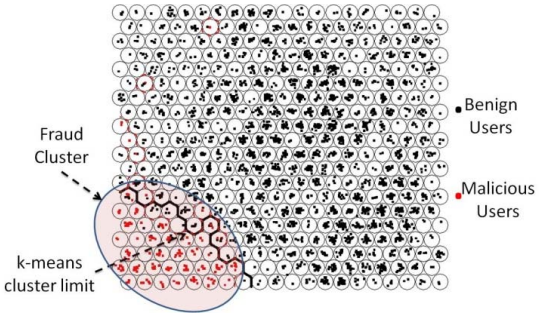

Figure 5: SOM/k-means clustering

VI. Simulation Results

In this section the two datasets presented in Chapter III will be used for large scale testing of the classifiers described in Chapter V. Classifier's parameters are arbitrarily chosen since their optimal values cannot anyhow be evaluated, due to the assumption of labeled samples absence. A sensitivity analysis has been conducted per classifier though in order to define parameter value regions for which the classifiers perform better, as far as F_1 score is concerned. The inspection number is always set to 5% of the consumers for classifiers that require a specific number of inspections. The impact of fraud intensity and fraud start time on each classifier's performance metrics is presented next.

A. CER data set-Residential consumers

In the first set of experiments the impact of residential consumers' fraud intensity is studied by fixing the fraud start time to day 165 and varying theft from 10%-90%. The experiments are executed 5 times simulating fraud in different consumers each time (same stands for paragraph B). The average values of metrics are presented in Fig. 6. The MGD classifier exhibits the best results as far as F_1 score and BDR are concerned. In addition, the Expert System ($\Delta\bar{E}_{min}$=0.5 and $\Delta\sigma_{min}$=0.2) performance is also high (second best). Clustering based methods (k-means, fuzzy c-means, SOM and DBSCAN/PCA) exhibit low performance metrics. The

number of clusters was set to 2 for k-means, fuzzy c-means and SOM. This choice highly impacts the results. Finally, the LOF/Rule classifier exhibits the worst results, since it is not able to recover large amounts of fraud even for high values of fraud intensity.

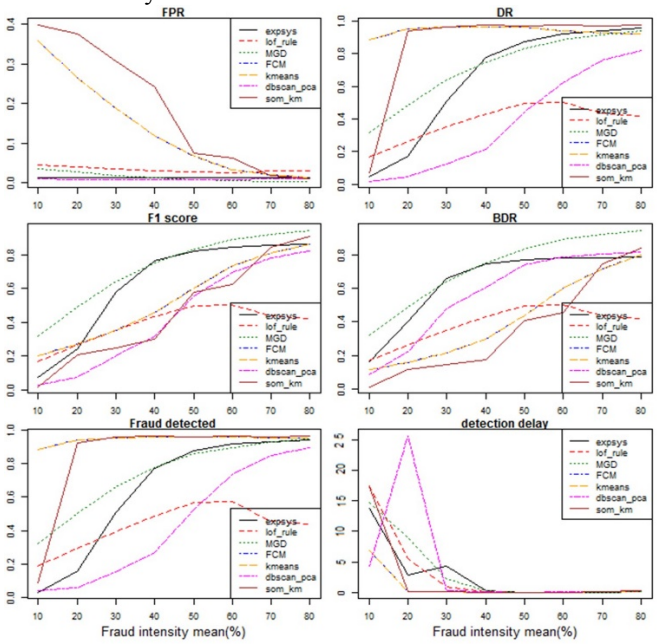

Figure 6: Classifier Performance metrics varying the fraud intensity for the CER data set residential consumers.

Next, the fraud start time is chosen randomly for each consumer between days 20-312 while fraud intensity is chosen from a beta distribution α=2, β=10. Experiments are executed 5 times and the average values of metrics are presented in Table I. The number of clusters for k-means and FCM is set to 4 this time. Overall the classifiers produce satisfactory results with LOF/Rule and DBSCAN performing worse. MGD and SOM give the best results as far as F_1 score and detected fraud are concerned. The expert system retrieves a large portion of stolen energy compared to FCM and k-means which exhibit similar behavior.

TABLE I. CER RESIDENTIAL CONSUMERS SCENARIO

Classifier	FPR (%)	DR (%)	F1 score (%)	BDR (%)	Fraud detected (%)	Detection Delay (days)
ExpSys	1.37	70.5	71.7	73.0	84.0	10
LOF/Rule	1.64	68.6	68.7	68.7	72.7	17.9
MGD	1.12	78.0	78.3	78.5	88.1	13.2
FCM	1.49	71.6	71.7	71.7	75.3	18.8
k-means	1.7	73.3	71.3	69.5	77.0	19.1
DBSCAN	1.03	66.2	71.3	77.2	70.8	17.6
SOM	1.37	76.0	75.5	76.4	87.0	13.7

The impact of fraud start time has also been studied and presented in Fig. 7. For this simulation fraud intensity receives values from a beta distribution (α=2, β=10) for each consumer. The data availability period (365 days) is divided in 12 even slots. In each simulation the fraud start parameter receives random values inside each slot thus producing 12 simulations. Each of these simulations are run 3 times with a different mixture of benign/malicious consumers and average metrics are calculated. Only F_1 score for the best classifier (MGD) is presented in Fig. 7, since the rest of the metrics for all the other classifiers follow the same trend. It can easily be concluded that committing fraud early (before day 50) or late (after day 250) in the data availability period significantly impacts the classifier's performance for residential consumers too (although the results are better than in the case of commercial consumers). In such cases F_1 scores are limited to 70% (the relative fraud detected metric falls below 75%).

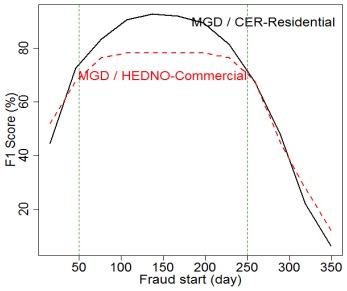

Figure 7: Impact of Fraud Start time on classification metrics

B. HEDNO data set-Commercial consumers

Commercial consumers have totally different load profiles than residential and thus are separately tested. In fact, commercial consumers' time series exhibit a lot of variability and sudden consumption drops which may be due to the nature of their business operation (for example not working on weekends, or working only in the summer). Such behavior may cause features to lose their discrimination capabilities and reduce classifiers' performance. Indeed, Fig. 8 suggests that all the classifiers perform worse than when applied on residential consumers. The MGD and expert system classifiers still exhibit the best results, although barely reaching 80% for F_1 score and BDR. Fuzzy c-means and SOM have been improved by choosing a number of clusters equal to 4 and inspecting the relative cluster centers for deciding which represent frauds. DBSCAN/PCA and LOF/Rule exhibit low performance metrics with k-means being the performing worst (not depicted in Fig. 8 for presentation reasons).

Next, the more realistic scenario where fraud start time is chosen randomly for each consumer between days 20-312 and fraud intensity is chosen from a beta distribution α=2, β=10 is demonstrated. Experiments are executed 20 times and the average values of metrics are presented in Table II. Simulation results indicate the superiority of MGD and the Expert System. Both achieve F_1 score close to 60% while maintaining FPR as low as 2.6% and detecting almost 75% of the stolen energy. SOM also produces satisfactory results, having the lowest FPR (1.53%) but its low DR (50.5%) is responsible for detecting only 64.4% of the stolen energy. In addition, comparing Table I with Table II (same scenarios on different data sets) reveals that the proposed approaches perform better in residential consumers than commercial.

The influence of fraud start time on F_1 score for the best classifier is presented in Fig. 7. Committing fraud early (before day 50) or late (after day 260) in the data availability period impacts the classifier's performance resulting to F_1 scores lower than 60% (the relative fraud detected metric falls below 50%). All performance metrics for all classifiers exhibit the same behavior.

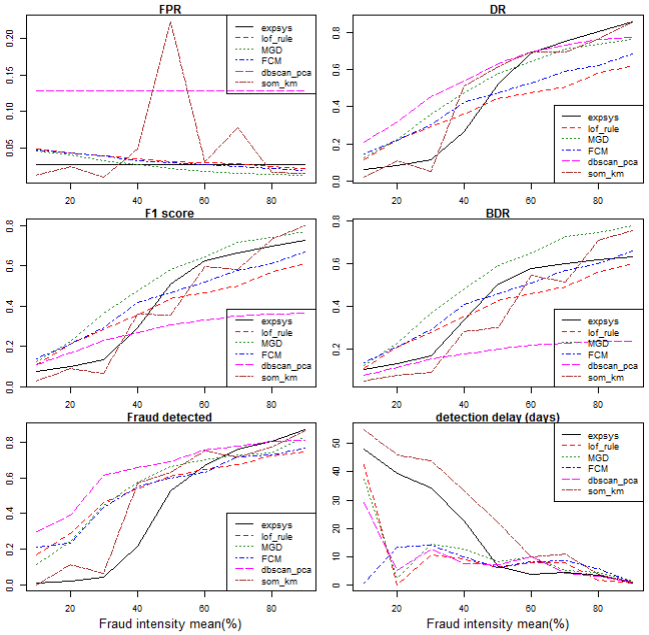

Figure 8: Classifier performance metrics varying the fraud intensity for the HEDNO data set commercial consumers

TABLE II. HEDNO COMMERCIAL CONSUMERS SCENARIO

Classifier	FPR (%)	DR (%)	F1 score (%)	BDR (%)	Fraud detected (%)	Detection Delay (days)
ExpSys	2.58	61.0	57.6	55.0	75.6	3.8
LOF/Rule	2.75	50.8	50.0	49.3	67.9	5.5
MGD	2.01	60.0	60.5	61.1	74.8	4.3
FCM	2.79	49.8	49.0	48.3	64.3	9.0
k-means	3.05	51.8	49.0	49.4	65.0	6.5
DBSCAN	12.8	62.7	30.6	20.3	75.6	8.5
SOM	1.53	50.5	53.3	59.7	64.4	5.5

VII. CONCLUSIONS

The authors tested different unsupervised learning and anomaly detection techniques for detecting NTL. The data sets used include commercial consumers from Greece and residential consumers from Ireland. The Twitter breakout detection library was used for the first time for extracting features that indicate NTL producing promising results. Seven unsupervised techniques were tested on both data sets studying the effect of fraud intensity and time fraud is initiated. The results are promising for both data sets, although they are better in the case of residential consumers. The MGD classifier and expert system exhibit higher performance in most cases detecting more than 85% (75%) of the stolen energy for residential (commercial) consumers when fraud intensity is more than 50%. The authors also demonstrate the impact of fraud initiation time, showing that frauds taking place early or late in the data availability period are harder to detect. This means that data availability periods must be extended in order to enhance detection. Finally, the proposed NTL detection schemes can accurately (error of 5-10 days for commercial and 10-20 days for residential consumers) detect the time fraud was initiated for confirmed malicious users.

ACKNOWLEDGMENT

The authors wish to thank V. Rogkakos, K. Andreadis and I. Menegatos of HEDNO for providing data used in this paper.

REFERENCES

[1] G. M. Messinis and N. D. Hatziargyriou, “Review of non-technical loss detection methods,” *Electr. Power Syst. Res.*, vol. 158, pp. 250–266, 2018

[2] P. Kadurek, J. Blom, J. F. G. Cobben, and W. L. Kling, “Theft detection and smart metering practices and expectations in the Netherlands,” in *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*, 2010, pp. 1–6.

[3] S. Weckx, C. Gonzalez, J. Tant, T. De Rybel, and J. Driesen, “Parameter identification of unknown radial grids for theft detection,” in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, 2012, pp. 1–6.

[4] S. A. Salinas and P. Li, “Privacy-Preserving Energy Theft Detection in Microgrids: A State Estimation Approach,” *IEEE Trans. Power Syst.*, pp. 1–12, 2015.

[5] 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 Deliv.*, vol. 26, no. 2, pp. 1284–1285, 2011.

[6] I. Monedero, F. Biscarri, C. León, J. I. Guerrero, J. Biscarri, and R. Millán, “Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees,” *Int. J. Electr. Power Energy Syst.*, vol. 34, no. 1, pp. 90–98, 2012.

[7] C. C. O. Ramos, A. N. De Sousa, 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, 2011.

[8] P. Jokar, N. Arianpoo, and V. C. M. Leung, “Electricity theft detection in AMI using customers’ consumption patterns,” *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[9] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, “Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid,” *IEEE Trans. Ind. Informatics*, vol. 12, no. 3, pp. 1005–1016, 2016.

[10] J. E. Cabral, J. O. P. Pinto, E. M. Martins, and A. M. A. C. Pinto, “Fraud detection in high voltage electricity consumers using data mining,” in *2008 IEEE/PES Transmission and Distribution Conference and Exposition*, 2008, pp. 1–5.

[11] E. W. S. Dos Angelos, O. R. Saavedra, O. a C. Cortés, and A. N. De Souza, “Detection and identification of abnormalities in customer consumptions in power distribution systems,” *IEEE Trans. Power Deliv.*, vol. 26, no. 4, pp. 2436–2442, 2011.

[12] V. Badrinath Krishna, G. A. Weaver, and W. H. Sanders, “PCA-Based Method for Detecting Integrity Attacks on Advanced Metering Infrastructure,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9259, no. Qest, 2015, pp. 70–85.

[13] J. I. Guerrero, C. León, I. Monedero, F. Biscarri, and J. Biscarri, “Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection,” *Knowledge-Based Syst.*, vol. 71, pp. 376–388, 2014.

[14] J. V Spirić, M. B. Dočić, and S. S. Stanković, “Fraud detection in registered electricity time series,” *Int. J. Electr. Power Energy Syst.*, vol. 71, no. 0, pp. 42–50, 2015.

[15] V. B. Krishna,, R. K. Iyer, and W. H. Sanders, "ARIMA-Based Modeling and Validation of Consumption Readings in Power Grids", vol. 9578, no. Critis. Cham: Springer International Publishing, 2016.

[16] D. Mashima and A. A. Cárdenas, “Evaluating Electricity Theft Detectors in Smart Grid Networks,” 2012, pp. 210–229.

[17] Irish Social Science Data Archive. [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/ .

[18] G. Messinis et al., “Utilizing smart meter data for electricity fraud detection,” CIGRE Science & Engineering Journal, June, 2017.

[19] S. Axelsson, “The base-rate fallacy and the difficulty of intrusion detection,” *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 3, pp. 186–205, Aug. 2000.

[20] BreakoutDetection R package. [Online] Available: https://github.com/twitter/BreakoutDetection. .

[21] N. A. James, A. Kejariwal, and D. S. Matteson, “Leveraging Cloud Data to Mitigate User Experience from ’ Breaking Bad ’,” pp. 3499–3508, 2016.