Anomaly Detection in Smart Grid Data: An Experience Report

Bruno Rossi, Stanislav Chren, Barbora Buhnova and Tomas Pitner
Faculty of Informatics
Masaryk University, Brno, Czech Republic
Email: {brossi,chren,buhnova,tomp}@mail.muni.cz

***Abstract*—In recent years, we have been witnessing profound transformation of energy distribution systems fueled by Information and Communication Technologies (ICT), towards the so called Smart Grid. However, while the Smart Grid design strategies have been studied by academia, only anecdotal guidance is provided to the industry with respect to increasing the level of grid intelligence. In this paper, we report on a successful project in assisting the industry in this way, via conducting a large anomaly-detection study on the data of one of the power distribution companies in the Czech Republic. In the study, we move away from the concept of single events identified as anomaly to the concept of collective anomaly, that is itemsets of events that may be anomalous based on their patterns of appearance. This can assist the operators of the distribution system in the transformation of their grid to a smarter grid. By analyzing Smart Meters data streams, we used frequent itemset mining and categorical clustering with clustering silhouette thresholding to detect anomalous behaviour. As the main result, we provided to stakeholders both a visual representation of the candidate anomalies and the identification of the top-10 anomalies for a subset of Smart Meters.**

***Index Terms*—Smart Grids, Smart Meters, Anomaly Detection, Clustering, Frequent Itemset Mining.**

I. Introduction

The Smart Grid can be regarded as an electricity network that benefits both from two-way cyber-secure communication technologies and computational intelligence for electricity generation, transmission, substations integration, distribution and consumption to reach the goals of a clean, safe, secure, reliable, resilient, efficient, and sustainable infrastructure [1].

The investment into large-scale Smart Grid deployment can be very risky, as confirmed for instance by investment losses during the Xcel Energys SmartGridCity project [1], [2].

There is a recent trend in adding more *"smartness"* in the Smart Grid infrastructure, so that the large amount of information that can be mined from normal usage can be used to drive the decision-making process and optimize the overall infrastructure management [3], [4]. This effect is enhanced by the two-way nature of the more modern infrastructures that allow operators to fine-tune parameters remotely based on the knowledge acquired from the operating conditions.

In the this paper, we deal with anomaly detection from Smart Grid data, that is looking for specific patterns in Smart Meter's data streams that do not conform to expected behaviour. In general terms, anomaly detection is a broad concept that has been applied to different fields, ranging from systems intrusion detection to fraud detection, with varying definitions of expected behaviour [5]. Based on real data from one of the power distribution companies in the Czech Republic, we propose an approach for the detection of the anomalies in the Smart Metering infrastructure that could be useful to promptly intervene to investigate the cause of unexpected behaviour. Based on this analysis, we report also about the insights acquired in terms of extensions of the approach that would allow us to implement such online system within the Smart Grid infrastructure.

The proposed approach is based on frequent itemset mining by encoding the different event types streamed from Smart Meters, applying segmentation of the itemsets and using categorical clustering for the evaluation of the itemsets and detection of unexpected patterns. The proposed approach is based on the analysis of event types from the Smart Meters. It allows us to detect anomalies that might have impact on the Smart Grid security, reliability or maintenance—for example suspicious manipulation with Smart Meter casing, under/over-voltage in specific locations or failure to switch remotely controlled appliances.

The paper is structured as follows. Section II overviews related work in the area of anomaly detection within Smart Grids. Section III then discusses the context of the study and provides descriptive information about the dataset. The anomaly detection approach is described in Section IV together with the rationale for its derivation. Section V presents the application within the Smart Grid domain according to the contextual information provided. The main evaluation and discussion from the experimental part is presented in Section VI, while Section VII brings up the conclusions.

II. Related Work

As the Smart Grid implementation is a strategic act for many countries, extensive attention has been paid to the study of smart infrastructures in recent years [1], [6]. Fang et al. [1] divide the smart infrastructure into three subsystems: (1) the *smart energy subsystem*, concerned with power generation, transmission, and distribution, (2) the *smart information subsystem*, concerned with information metering, measurement, and management, and (3) the *smart communication subsystem*,

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

concerned with wireless and wired communication technologies, and the end-to-end communication management.

Within the *smart information subsystem* (2), which frames our work, significant advancement can be observed in both industry and academia.

In industry, these projects are mainly led by electric utilities or related organizations, which are however often lacking expertise in information and communication technologies [1]. Evaluation of the devised strategies is hence realised rather via pilot projects than via analytical and simulation means. While simulation is not that uncommon within the design of smart energy subsystem (1) and smart communication subsystem (3), it is rather rare within the smart information subsystem (2) [7].

In academia, many approaches for the analysis of data flowing within the Smart Grid, and hence the identification of smart information, exist, mostly in the cyber-security domain [8], [9]. Within the cyber-security, the approaches are mainly concerned with intrusion detection harming the confidentiality, integrity, or availability of the Smart Grid [10], [11], [12], although more work has been done on preventative measures, such as secure communication protocols and architectures in the Smart Grid [8]. Overall, the cyber-security in the frame of Smart Grids is very well researched and hence we will invest more effort in investigating the other domains.

Besides cyber-security, the analysis of the Smart Grid information flow is concerned mainly with the detection of faults and failures [13], [14], [15], and to minor extent with the study of consumer behaviour [16]. Calderaro et al. [13] detect failures in data transmission and faults in the distribution network with the help of petri nets analysis and matrix operations. Kalaitzis et al. [14] study powerline faults (on the level of the amplitude, frequency, or phase of powerline signal) with the help of a sliding window approach in multivariate time series. He et al. [15] are concerned with fault detection and localization in transmission lines, using a network inference algorithm based on Markov random fields and dependency graphs. However, all these approaches are based on the powerline level (i.e. modelling and observing individual relays [13], powerline signal [14] or phasor angles across the transmission line buses [15]), not the information flow above it, which differentiates them from our aim.

The application of clustering to Smart Grids data is not a novel idea, and has been successfully applied to Smart Grid data to steer towards a more intelligent Smart Grid infrastructure [3], [4]. However, applications are more specific to clustering customers according to behaviour from Smart Meters data [17] or looking into clustering sensor data to segment the network topology and identify set of clusters according to energy profiles [18].

III. Context of the Study

This study was conducted in cooperation with the major energy distribution company in the Czech Republic, in which the smart metering infrastructure has been tested and examined in several pilot projects since 2006. The pilot projects have been part of the European Grid4EU initiative. Currently, there have been almost 40,000 Smart Meters deployed in total which constitutes about 1% of all consumers managed by the Distribution company. The individual pilot projects have specific goals, e.g. evaluation of available technologies, communication infrastructure, quality of service.

In our case, we utilise the data sets from a project focusing on local load management in low voltage power grids. The selected consumers (both households and industry) are equipped with Smart Meters that collect data about the power consumption profile. The data is periodically sent to the data concentrator which is installed at the Distribution Transformation Station (DTS)—there is one data concentrator per each DTS. From the data concentrator, the data is collected and stored by the Data Central (DC) server, which is located at the power grid operation centre. Besides the data related to the power consumption attributes, the individual devices generate variety of events used for the grid monitoring and maintenance.

A central role is played by the Smart Meter, an electronic device used for the measurement and provision of billing information to customers [19]. In this study, we consider Smart Meters as data sources of data streaming, allowing the analysis of all data derived from the Smart Meters' operations.

The Smart Meter events are used to notify the data central about important state changes that happened at the level of the Smart Meter, such as powering up or down of the meter, tariff and rate switching, time synchronization, etc... Each event belongs to one of possible 76 event types.

An event entry is described by the origin time, event type and a Smart Meter device it was created by. For the device, there is a number of additional attributes available, such as its GPS coordinates, date of installation, tariff category or type of deployment site (e.g. apartment, house, agriculture/industry, etc.). The entire dataset contains 364,107 event entries that have been collected from 381 Smart Meters over the time-span from December 2014 to July 2015.

IV. Anomaly Detection Approach Definition

When dealing with anomaly detection, one important distinction is based on the way we aggregate data to determine unexpected behaviour. We usually distinguish among: i) point anomalies, ii) context anomalies, and iii) collective anomalies [5]. A point anomaly means that one individual event instance can be considered anomalous when compared to the remaining data. For example, counting the number of occurrences of a *"gateway on"* event from a Smart Meter might be considered anomalous if its frequency is too low or high on a specific day. Context anomalies start from the assumption of dividing the behaviour from the context: the same behaviour might not be considered an anomaly if it happens in a different context. Based on the previous example, the same number of occurrences of *"gateway on"* might not trigger an anomaly detection mechanism if they happen on a specific time of the day / period of the year. Compared to point anomaly, we need to take into account the context of the event instance. The third—and more interesting for our context—category is referred to as collective anomaly. In this case, the event instance does

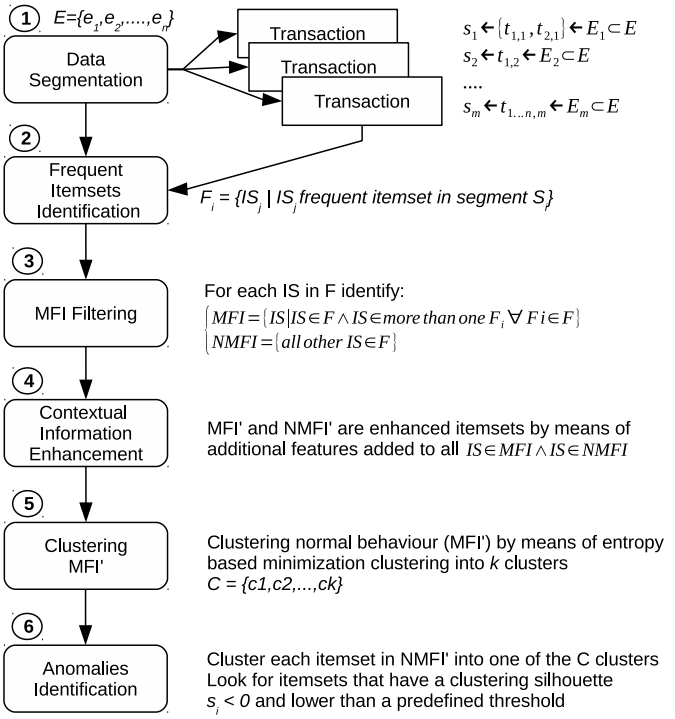

Fig. 1. Proposed Approach for Smart Meters anomaly detection (E=set of events, S=set of data segments, T=set of transactions, F=frequent itemset, MFI=most frequent itemset, C=set of clusters).

not represent an anomaly *per se*, but only if considered within the collection of all the other events instances. Continuing the previous example, we might consider as anomalous the collection of events *"gateway on"*,*"gateway off"*,*"gateway on"*,*"gateway off"* rather than *"gateway on"*,*"transmission start"*,*"gateway off"*. For this type of anomaly, looking at single event instances is not meaningful, as they need to be considered together with the other collection of events.

A typical further characterization of anomaly detection is based on the availability of label data for the event instances that can constitute anomalies (supervised, semi-supervised and unsupervised approaches [5]). In the former category, there is the availability of labelled data for either all instances or positive instances, meaning usually one human expert dealing with labelling of each anomalous instance or the availability of some form of failure data from which labels can be derived—also known as tagging information. In our specific context, we did not have any of such information available, so we ruled-out the option of applying a supervised classification approach for anomaly detection. In fact, due to the characteristics of the Smart Metering dataset we opted for what we can refer to as an *unsupervised contextual and collective anomaly detection* approach. The main reasons were the unavailability of tagging information, the needs to consider events within their context and within the broader concept of itemsets.

Given all the considerations about the way to tackle the problem in the Smart Grid domain, our approach is similar to the one of Barbará et al. that has been successfully applied in the context of intrusion detection [20]. However, in our approach we use a different way to detect outliers — clustering silhouette indicators. Furthermore, by applying the approach to Smart Grids data streams we identified several improvements that we discuss in the paper. The approach is based on the idea to first identify clusters of what can be considered as normal behaviour, and then to look into itemsets that deviate from the knowledge learned from the dataset. We present in detail the steps of the approach (Fig. 1):

Step1. We first apply Association Rule Mining to identify frequent itemsets [21], that is sets of events instances that are more recurrent. For this, we need to identify sets of transactions from the data streams. We define one transaction as the set of all the items derived per one day and per each Smart Meter (*Data Segmentation*, Fig. 1, *Step1*). Thus, each Smart Meter will be associated with a list of daily transactions of operations;

Step2. Based on the aforementioned concept of *collective anomaly*, we extract the most frequent itemsets from data transactions by applying the the Apriori algorithm [21]. This will yield for each Transaction a list of frequent itemsets within each transaction (*Frequent Itemsets Identification*, Fig. 1, *Step2*). As an example, after running the first two steps we might end up with the following itemsets:

```
{R2XR1On,Rate 2 switching}
{Overcurrent L1,Overcurrent L3}
{R2XR1Off,R2XR1On}
{R2XR1Off,R2XR1On,Rate 1 switching}
[...]
```

Step3. For each data segment, we have now the list of frequent itemsets derived from all the transactions. We look then for the Most Frequent Itemsets that are present in more than one segment (*MFI Filtering*, Fig. 1, *Step3*). The assumption is that the itemsets that appear in more than one segment can be considered as an initial normal behaviour, while the other can be considered *potential* anomalies at this stage;

Step4. Following the concept of *contextual anomaly*, each frequent itemset is further augmented with additional information, so that the same itemset in different segments will be represented by additional features, e.g. whether a *working day* (*Contextual Information Enhancement*, Fig. 1, *Step4*). Note that this will increase the number of features, but will also increase the number of itemsets as the same itemset might appear in two different contexts (*working day / non-working day*). At this point, the additional features can be derived from additional data sources, not only Smart Meters—as long as they can be associated to the itemsets. An example of itemsets at this stage:

```
{R2XR1On,Rate 2 switching,week-day}
{Overcurrent L1,Overcurrent L3,week-day}
{Overcurrent L1,Overcurrent L3,week-end}
[...]
```

Step5. We cluster all the *normal* itemsets identified in Step2. The assumption is that we cluster these as representative of the normal behaviour (*Clustering MFI'*, Fig. 1, *Step5*). We might also look at this stage from the clustered data if there are

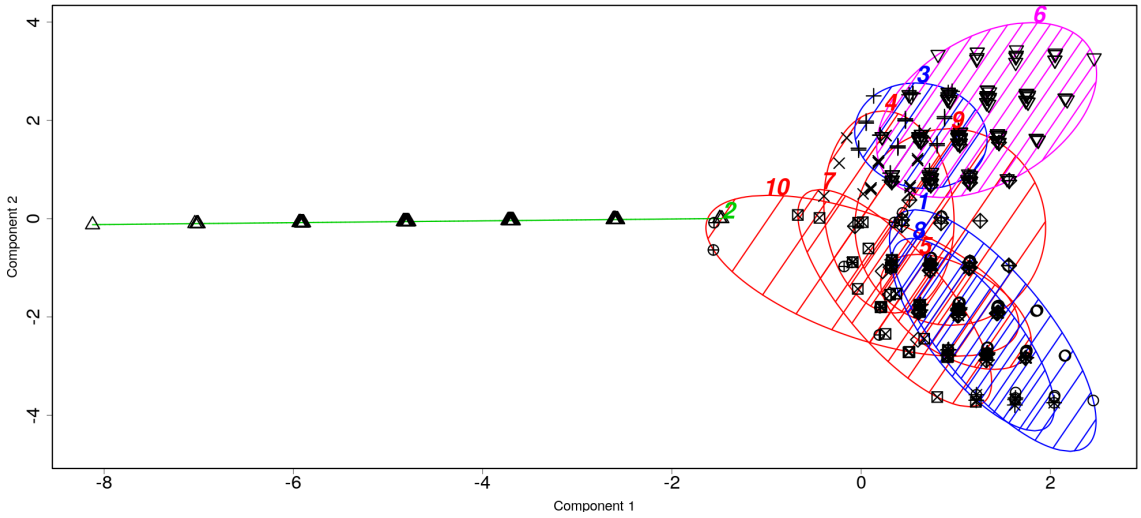

Fig. 2. Clustering of 17 devices (n=681, k=10)

itemsets that are isolated or do not fit well in their clusters. As the itemsets are represented by categorical / nominal variables, we use a categorical clustering based on entropy minimization [22] — used also in the original approach. To continue with the example, if clustering the previous data with a number of clusters $k = 2$, we can get the following clusters as they minimize the entropy:

`C1: {R2XR1On,Rate 2 switching,week-day}`
`C2: {Overcurrent L1,Overcurrent L3,week-day}`
`C2: {Overcurrent L1,Overcurrent L3,week-end}`

Step6. The final step looks into the identification of anomalies (*Anomalies Identification*, Fig. 1, *Step6*). In this step, we consider again all the itemsets that were considered as *potential* anomalies in Step2. For each of them, we cluster them and we look how well they fit according to the clusters created with the previously clustered data.

For this last step, to identify the goodness of fit of the new itemsets, we use the concept of *clustering silhouette* [23]. Given an itemset i, $a(i)$ represents the average dissimilarity between all the other itemsets in i's cluster. Given all the other clusters ($\forall C_j \, where \, i \notin C_j$), $d(i, Cj)$ represents the dissimilarity of i with all the itemsets in C_j and $b(i) = min(d(i, Cj))$ represents the minimal distance of itemset i to the nearest cluster. The silhouette represents how well an element fits in its cluster:

$$s_i = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

We can have three cases:

- $s_i > 0$: the itemset fits generally well into the cluster;
- $s_i \approx 0$: the itemset is clustered between two clusters;
- $s_i < 0$: the itemset is probably clustered in the wrong cluster, that is from the silhouette definition, the itemset has higher dissimilarity with the elements of the belonging cluster than some nearby cluster;

When looking for anomalies we look for the third type of cases, that is those in which the itemset does not fit well in created clusters. Furthermore, we can set a threshold value for the silhouette. In the next section we present the application of the approach to the Smart Meter data and we will also use a visual representation of the *clustering silhouette* to aid in anomalies identification.

V. Application within the Smart Grids Domain

The first consideration in analysing the Smart Meter data is about parameters fine-tuning. Overall, for the frequent itemset mining we need to define *support*—proportion of transactions that contain a specific itemset—and *confidence*—for association rules $X \implies Y$, the proportion of the transactions that contain both X and Y. In the current analysis we used *support*=0.1 and *confidence*=0.8. Varying these parameters can bring a different number of itemsets considered in the initial steps. A third relevant parameter is the number of clusters, k. Unluckily, the identification of the best number of clusters based on the underlying dataset can be computationally demanding and unfeasible for a large number of Smart Meters. Sensitivity analysis can be used to optimize some clustering quality indicator but such approach does not scale up to larger number of itemsets. In the current experimental section, we used $k = 10$. A last parameter is the threshold for the *clustering silhoutte* to determine the anomalies. We set this parameter to -0.20, but such value does not need to be evaluated apriori, and can be set by looking at the visual representation of the *clustering silhouette*.

Running the approach on the dataset with 381 devices brings to a total of 364,107 overall events generated on the devices based on the 76 event types. We map these events into 20,670 transactions associated to the 381 segments (*Step1*). Running the Apriori algorithm brings overall 273,829 non-unique itemsets (*Step2*). We can note how this is a large

number of itemsets, due to the fact that at this stage the same itemset might appear in different transactions. The next step goes into filtering the itemsets based on the condition that they are present in ≥ 2 segments considering the unique itemsets at this stage, not anymore their association to a transaction. This brings 44,450 unique itemsets that can be considered normal behaviour according to the discussion above (*Step3*).

The list of itemsets at this stage is too large to give useful insights to a decision maker. We can at this stage add more contextual information (*Step4*). Since we are considering unique events in *Step3*, by mapping back the events to the ones in the segment we increase the number of events, as an event might happen under a different context. This differentiates further two events and is conforming to the idea of augmenting the data points with contextual information. In the current analysis we skipped this step, but it is applicable with the assumption that the added contextual information is categorical or can be converted in such form.

We run then the clustering algorithm on the mapped events to create clusters of itemsets that are similar in terms of the identified categorical features (*Step5*). We use categorical clustering based on entropy minimization [22]. To represent the clusters, we use the *Clusplot* representation, in which clusters are represented as ellipses after multi-dimensionality reduction by showing the two principal components [24].

After clustering has been completed, we have a set of clusters that represent the way in which all the itemsets are mapped to each cluster. To simplify the representation, we show the categorical clustering with $k = 10$ clusters performed on 681 itemsets for 17 devices as a subset of the whole dataset (Fig. 2). The size of the clusters shows the spread of the itemsets within the cluster and the shadowed area shows the density within clusters. Clustered data represent the normal behaviour according to our initial assumptions. However, as noted in the previous section, we can still see areas worth investigation, like cluster nr. 2 that is more spread apart compared to the others.

We run then the final step (*Step6*). We now take back the itemsets that were not frequent in the original dataset (*NMFI'*). We cluster each one of the new itemsets to see how well they fit into the existing clusters (based on *MFI'*).

If we look at the fitting of different itemsets for the case of 17 devices, we can represent each itemset according to the *silhouette width* (Fig. 3). We can see that some itemsets on the upper part of the plot (s_i near 1.0) are fitting better in the clusters. On the bottom part, those that have negative s_i. As new itemsets are clustered according to Step6 of the approach, we look for those that have negative s_i.

Based on the analysed data, if we set a threshold of -0.20, we can identify the *top-10 anomalies* after running the overall approach (Table I). We report events that were detected as anomalies together with their s_i *silhouette* value and a reference index i. In particular, there are five itemsets that might signal some forms of *missing voltages* associated with *overcurrents* (itemset indexed $1, 3, 4, 8, 9$). These are itemsets that can be worth further investigation by decision makers.

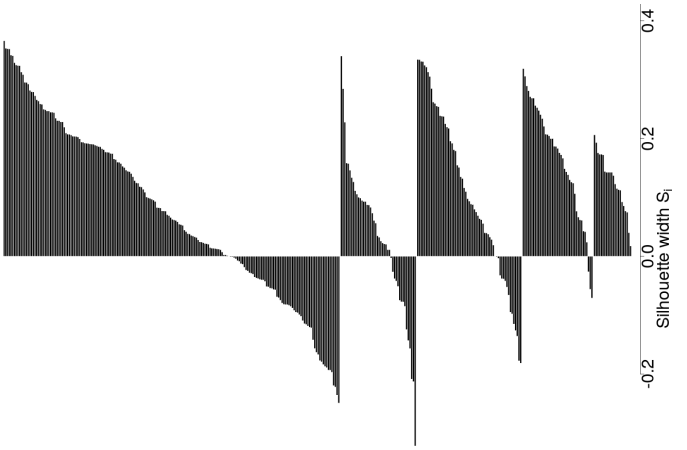

Fig. 3. Lower part of the Clustering Silhouette, 17 devices (n=681, k=10)

VI. Evaluation & Discussion

There are several lessons learned from the application of the approach to anomaly detection in Smart Meters data.

We approached initially the problem from different angles but we found out that the most important aspect when considering anomalies is to provide a *collective* and *contextual* overview. At least in the experimental project we described, *single point anomaly detection* did not prove to be sufficient to determine anomalous events. The addition of contextual information and the inclusion of an event type instance in relation to other event types instances proved to be a more powerful mechanism.

TABLE I
Top-10 Anomalies according to s_i, devices=17, n=681, k=10

s_i	i	itemset
-0.322	1	{"Missing voltage L2", "Overcurrent L1", "Overcurrent L2", "Overcurrent L3"}
-0.255	2	{"Limiter activated", "Power-up"}
-0.249	3	{"Limiter activated", "Overcurrent L1", "Power-down" "Power-up", "Rate switching error clear Rate switching error cleared in meter"}
-0.239	4	{"Missing voltage L2", "Overcurrent L2"}
-0.236	5	{"Limiter activated", "Overcurrent L1", "Power-down", "Rate switching error"}
-0.222	6	{"Limiter activated", "Overcurrent L1", "Power-down", "Rate switching error clear Rate switching error cleared in meter"}
-0.216	7	{"Rate 1 switching", "Rate 2 switching"}
-0.212	8	{"Missing voltage L2", "Overcurrent L2", "Overcurrent L3"}
-0.219	9	{"Overcurrent L1", "Power-down", "Power-up", "TOU activated meter"}
-0.208	10	{"No overcurrent L2 ...", "Overcurrent L1", "Overcurrent L3"}

An additional aspect in the Smart Grids domain is that—differently from other domains—we are not aware of existing datasets that can be used for the evaluation of the goodness of anomaly detection from Smart Meters data in comparison with

an established ground truth. As such, the opinion of domain experts becomes very relevant: it is however unrealistic to provide indicators such as *false negatives* and *false positives* due the vast amount of itemsets to review. This makes the evaluation of the approaches difficult to perform.

Running the project also allowed to identify several drawbacks of the approach. These constitute an interesting list of requirements for the improvement of the implemented solution. One of the drawbacks of the proposed approach is that we considered itemsets and not sequences. That is, we did not discriminate the order of events within itemsets both in frequent itemset mining and in clustering. One of the aspects we are keen to explore is the usage of sequences for the segmentation part with a different algorithm to cluster sequences. Together with experts opinions we might derive a comparison of several approaches.

Another consideration is about using the approach for online learning—streaming data and real-time system behaviour. This poses different issues than those considered in this paper, but working towards the implementation of such a system can be useful to support the concept of "smarter grids".

Finally, in this paper we did not explore in detail the usage of contextual information in the experimental part. Given how the features have been built this is not a problem as long as numerical features are converted to categorical data. We were considering also an initial phase in which domain experts could rule-out non-interesting or non-relevant event types. However, this initial phase can also be detrimental to the possibility to detect unexpected events.

VII. Conclusion

Modern Smart Grids will permeate our lives in years to come. While in their initial appearances data communication was mostly one-way, we are now in the context of two-ways Smart Grids that can not only monitor but also fine-tune behaviour based on knowledge mining capabilities. In this sense, there is a growing need to introduce smarter behaviours in the infrastructure, and a central role is played by Smart Meters as devices that can engage in two-way communications.

In this paper, we evaluated an approach for anomaly detection in Smart Grids derived from data streamed from Smart Meters. We proposed to approach the problem by taking into account the aspects of *collective* and *contextual* anomalies that can bring benefits in building a wider set of dependencies among events derived from Smart Meters.

We presented the application of a proposed *unsupervised contextual and collective detection approach* to data streams from a large energy distributor in the Czech Republic to reason about different types of possible anomalies (e.g. *over-voltages*, *under-voltages*). We discussed the benefits of the approach but also identified drawbacks that can lead to improvements towards the implementation of an online learning system.

In running the project we found several key needed characteristics: the necessity to provide a constantly online-learning system, scalable, and that can support detection of unexpected events, possibly leading towards a self-healing system.

References

[1] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid — the new and improved power grid: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 14, no. 4, pp. 944–980, 2012.

[2] "Xcel energy. smartgridcity," http://www.xcelenergy.com/.

[3] C.-W. Tsai, A. Pelov, M.-C. Chiang, C.-S. Yang, and T.-P. Hong, "A brief introduction to classification for smart grid," in *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, ser. SMC '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 2905–2909.

[4] C. S. Lai and L. L. Lai, "Application of big data in smart grid," in *2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2015, pp. 665–670.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[6] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G. P. Hancke, "A survey on smart grid potential applications and communication requirements," *Industrial Informatics, IEEE Trans. on*, vol. 9, no. 1, pp. 28–42, 2013.

[7] K. Mets, J. A. Ojea, and C. Develder, "Combining power and communication network simulation for cost-effective smart grid analysis," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 3, pp. 1771–1796, 2014.

[8] W. Wang and Z. Lu, "Cyber security in the smart grid: Survey and challenges," *Computer Networks*, vol. 57, no. 5, pp. 1344–1371, 2013.

[9] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on cyber security for smart grid communications," *Communications Surveys & Tutorials, IEEE*, vol. 14, no. 4, pp. 998–1010, 2012.

[10] Y. Zhang, L. Wang, W. Sun, R. C. Green, M. Alam *et al.*, "Distributed intrusion detection system in a multi-layer network architecture of smart grids," *Smart Grid, IEEE Trans. on*, vol. 2, no. 4, pp. 796–808, 2011.

[11] R. Berthier, W. H. Sanders, and H. Khurana, "Intrusion detection for advanced metering infrastructures: Requirements and architectural directions," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*. IEEE, 2010, pp. 350–355.

[12] C.-W. Ten, J. Hong, and C.-C. Liu, "Anomaly detection for cybersecurity of the substations," *Smart Grid, IEEE Trans. on*, vol. 2, no. 4, pp. 865–873, 2011.

[13] V. Calderaro, C. N. Hadjicostis, A. Piccolo, and P. Siano, "Failure identification in smart grids based on petri net modeling," *Industrial Electronics, IEEE Trans. on*, vol. 58, no. 10, pp. 4613–4623, 2011.

[14] A. Kalaitzis and J. D. Nelson, "Online joint classification and anomaly detection via sparse coding," in *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*. IEEE, 2014, pp. 1–6.

[15] M. He and J. Zhang, "A dependency graph approach for fault detection and localization towards secure smart grid," *Smart Grid, IEEE Trans. on*, vol. 2, no. 2, pp. 342–351, 2011.

[16] S. Lühr, G. West, and S. Venkatesh, "Recognition of emergent human behaviour in a smart home: A data mining approach," *Pervasive and Mobile Computing*, vol. 3, no. 2, pp. 95–116, 2007.

[17] M. Zeifman, "Smart meter data analytics: Prediction of enrollment in residential energy efficiency programs," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2014, pp. 413–416.

[18] P. P. Rodrigues and J. Gama, "Holistic distributed stream clustering for smart grids," in *Workshop on Ubiquitous Data Mining*, 2012, p. 18.

[19] "Meters, smart. "smart meter systems: a metering industry perspective." a joint project of the eei and aeic meter committees (2011)."

[20] D. Barbará, Y. Li, J. Couto, J.-L. Lin, and S. Jajodia, "Bootstrapping a data mining intrusion detection system," in *Proc. of the 2003 ACM symposium on Applied computing*. ACM, 2003, pp. 421–425.

[21] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *20th international conference on very large data bases*. Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.

[22] D. Barbará, Y. Li, and J. Couto, "Coolcat: an entropy-based algorithm for categorical clustering," in *Proc. of the eleventh international conference on Information and knowledge management*. ACM, 2002, pp. 582–589.

[23] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[24] G. Pison, A. Struyf, and P. J. Rousseeuw, "Displaying a clustering with clusplot," *Computational statistics & data analysis*, vol. 30, no. 4, pp. 381–392, 1999.